Unable to locate subtitle

AWS Well-Architected Framework

AWS Well-Architected Framework: ***Unable to locate subtitle***

Table of Contents

AWS Well-Architected Framework

Data de publicação: 20 de outubro de 2022 (Revisões do documento)

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.

Introdução

O AWS Well-Architected Framework ajuda a entender os prós e os contras das decisões que você toma ao criar sistemas na AWS. O uso do Framework ajuda você a aprender as práticas recomendadas de arquitetura para projetar e operar workloads confiáveis, seguras, eficientes, econômicas e sustentáveis na Nuvem AWS. 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 sistemas bem projetados aumenta significativamente a probabilidade de êxito dos negócios.

Os arquitetos de soluções da AWS têm vários anos de experiência em arquitetura de soluções em uma ampla variedade de segmentos de negócios verticais 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 práticas recomendadas 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. À medida que a AWS evoluir, e continuarmos a aprender mais com o trabalho com nossos clientes, aprimoraremos ainda mais 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 práticas recomendadas e as estratégias da AWS a serem usadas ao projetar e operar uma workload na nuvem e fornece links para detalhes de implementação e padrões de arquitetura adicionais. Para mais informações, leia a Página inicial do AWS Well-Architected.

A AWS também fornece um serviço para analisar suas workloads gratuitamente. A Ferramenta AWS Well-Architected (Ferramenta AWS WA) é um serviço na nuvem que fornece um processo consistente para você analisar e medir sua arquitetura usando o AWS Well-Architected Framework. A Ferramenta AWS WA fornece recomendações para tornar suas workloads mais confiáveis, seguras, eficientes e econômicas.

Para ajudá-lo a aplicar as melhores práticas, criamos os AWS Well-Architected Labs, que fornecem um repositório de código e documentação para oferecer experiência prática na implementação das melhores práticas. Também nos associamos a parceiros selecionados da Rede de Parceiros da AWS (APN), que são membros do Programa de parceiros do AWS Well-Architected. Esses parceiros da AWS têm profundo conhecimento sobre a AWS e podem ajudar você a analisar e melhorar suas workloads.

Definições

Todos os dias, os especialistas da AWS ajudam os clientes a projetar sistemas para aproveitar as práticas recomendadas 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 práticas recomendadas para os clientes e parceiros avaliarem arquiteturas e um conjunto de perguntas que você pode usar para avaliar o alinhamento de uma arquitetura com as práticas recomendadas da AWS.

O AWS Well-Architected Framework é baseado em seis pilares: Excelência operacional, Segurança, Confiabilidade, Eficiência de performance, Otimização de custos e Sustentabilidade.

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.

Nome	Descrição
Segurança	O pilar de segurança descreve como aproveitar as tecnologias de nuvem para proteger dados, sistemas e ativos de uma maneira que possa melhorar sua postura de segurança.
Confiabilidade	O pilar Confiabilidade abrange a capacidade de uma carga de trabalho de executar a função pretendida correta e consistentemente quando esperado. Isso inclui a capacidade de operar e testar a carga de trabalho durante todo o ciclo de vida dela. Este documento fornece orientaçõ es detalhadas sobre as práticas recomendadas para a implementação de workloads confiáveis na AWS.
Eficiência de performance	A capacidade de usar recursos de computaçã o 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.
Sustentabilidade	O pilar Sustentabilidade inclui a possibilidade de melhorar continuamente os impactos sobre a sustentabilidade com a redução do consumo de energia e o aumento da eficiência de todos os componentes de uma workload por meio da maximização dos benefícios dos recursos provisionados e da minimização do total de recursos necessários.

No AWS Well-Architected Framework, usamos estes termos:

- A 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 dissociada de outros componentes.
- O termo carga de trabalho é usado para identificar um conjunto de componentes que entrega o valor empresarial. Uma carga de trabalho é normalmente o nível de detalhes sobre o qual os líderes de negócios e tecnologia se comunicam.
- Pensamos na arquitetura como sendo os componentes que trabalham juntos em uma carga de trabalho. Como os componentes se comunicam e interagem é, com frequência, o foco dos diagramas de arquitetura.
- Marcos assinalam as principais alterações em sua arquitetura, à medida que evoluem ao longo do ciclo de vida do produto (design, implementação teste, ativação e produção).
- Dentro de uma organização o portfólio de tecnologia é a coleção de cargas de trabalho necessárias para o negócio operar.
- O nível de esforço refere-se à categorização da quantidade de tempo, esforço e complexidade que uma tarefa exige para implementação. Cada organização precisa considerar o tamanho e a especialização da equipe e a complexidade da workload a fim de ter contexto adicional para categorizar adequadamente o respectivo nível de esforço.
  - Alto: O trabalho pode levar várias semanas ou vários meses. Isso poderia ser dividido em vários lançamentos, histórias e tarefas.
  - Médio: O trabalho pode levar vários dias ou várias semanas. Isso poderia ser dividido em vários lançamentos e tarefas.
  - Baixo: O trabalho pode levar várias horas ou vários dias. Isso poderia ser dividido em várias tarefas.

Ao arquitetar cargas de trabalho, você obtém vantagens e desvantagens entre os pilares com base no contexto da sua empresa. Essas decisões de negócios podem definir suas prioridades de engenharia. Você pode otimizar para melhorar o impacto sobre a sustentabilidade e reduzir os custos à custa da confiabilidade em ambientes de desenvolvimento ou, no caso de soluções de missão crítica, otimizar a confiabilidade e aumentar os custos e o impacto sobre a sustentabilidade. 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 incluem um conjunto de funções, como arquiteto técnico (infraestrutura), arquiteto de soluções (software), arquiteto de dados, arquiteto de redes e arquiteto de segurança. Geralmente, essas equipem usam o TOGAF ou o Zachman Framework como parte de um recurso de arquitetura corporativa.

Na AWS, preferimos distribuir os recursos para as equipes, em vez de termos uma equipe centralizada com esses recursos. 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, nós temos práticas (processos, padrões, normas aceitas e formas de fazer as coisas) que são destinados a permitir que cada equipe tenha essa capacidade, e utilizamos especialistas que garantem que as equipes elevem o nível dos padrões que elas precisam cumprir. Segundo, implementamos mecanismos que realizam verificações automatizadas para garantir que os padrões sejam atendidos.

> “Boas intenções nunca funcionam, você precisa de bons mecanismos para fazer qualquer coisa acontecer” — Jeff Bezos.

Isso significa substituir os melhores esforços humanos por mecanismos (muitas vezes automatizados) que examinam a conformidade com base em regras ou processos. Essa abordagem distribuída é embasada pelos princípios de liderança da Amazone estabelece uma cultura em todas as funções que retornam do cliente. Trabalhar de trás para a frente é uma parte fundamental do nosso processo de inovação. Começamos com o cliente e o que ele quer, e deixamos isso definir e orientar nossos esforços. 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 obter essas capacidades ou as equipes existentes a elevar seus padrões, viabilizamos o acesso a uma comunidade virtual de engenheiros-chefes que podem analisar os projetos e ajudá-las a entender quais são as práticas recomendadas 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 práticas recomendadas 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, como engenheiros-chefes, para defini-los. À medida que os engenheiros-chefes veem surgir novas melhores práticas, eles trabalham como uma comunidade para garantir que elas sejam seguidas pelas equipes. 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 Framework é a implementação voltada para o cliente do nosso processo de análise interna, no qual codificamos nosso pensamento de engenharia principal nas funções de campo, como a arquitetura de soluções e equipes de engenharia internas. O Well-Architected Framework é um mecanismo escalável que permite que você aproveite esses aprendizados.

Seguindo a abordagem de uma comunidade de engenheiros-chefes com propriedade distribuída de 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:

- Pare de adivinhar suas demandas de capacidade: se você tomar uma decisão ruim relacionada à capacidade ao implantar uma carga de trabalho, poderá acabar com recursos ociosos caros ou lidando com as implicações da performance da capacidade limitada. 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 suas cargas de trabalho a baixo custo e evitar a despesa de esforços manuais. Você pode acompanhar as alterações em sua automação, auditar o impacto e reverter para os parâmetros anteriores, quando necessário.
- 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 evoluir, 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 de arquitetura 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 pilares do Framework

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

Pilares

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

Excelência operacional

O pilar Excelência operacional inclui a capacidade de oferecer suporte ao desenvolvimento e de executar cargas de trabalho com eficácia, obter insights sobre as operações e melhorar continuamente processos e procedimentos de suporte para oferecer valor empresarial.

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

Tópicos

- Princípios de design
- Definição
- Práticas recomendadas
- Recursos

Princípios de design

Existem cinco princípios de design para excelência operacional na nuvem:

- Execute 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 workload (aplicações, 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.
- Faça alterações frequentes, pequenas e reversíveis: projete 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).
- Refine os procedimentos operacionais com frequência: ao usar 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.
- Antecipe falhas: execute 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 melhorias com as lições aprendidas em todos os eventos e falhas operacionais. Compartilhe o que foi aprendido com as equipes e a organização inteira.

Definição

Existem quatro áreas de melhores práticas para excelência operacional 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.

Práticas recomendadas

Tópicos

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 na 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, alterações 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, a fim de minimizar o impacto, e os incentive a engajar os tomadores de decisão e as partes interessadas quando acharem que há algum risco, para resolvê-lo 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 normativos ou de conformidade externos aplicáveis à sua organização, você deverá usar os recursos fornecidos pela Conformidade com a Nuvem AWS para ajudar a educar suas equipes, para que elas possam determinar o impacto sobre suas prioridades. O Well-Architected Framework enfatiza o aprendizado, a medição e a melhoria. Ele oferece uma abordagem consistente para avaliar arquiteturas e implementar designs que escalem ao longo do tempo. A AWS fornece o AWS Well-Architected Tool para ajudar você a analisar sua abordagem antes do desenvolvimento e o estado de suas workloads antes da produção e durante a produção. Você pode comparar as workloads com as práticas recomendadas de arquitetura da AWS mais recentes, monitorar seu status geral e obter insights sobre possíveis riscos. O AWS Trusted Advisor é uma ferramenta que fornece acesso a um conjunto essencial de verificações que recomendam otimizações capazes de 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 suas prioridades.

A AWS pode ajudar a instruir suas equipes sobre a AWS e os serviços que ela fornece para que compreendam melhor como as escolhas que elas fazem podem ter um impacto na workload. Use os recursos fornecidos pelo AWS Support (Centro de Conhecimento da AWS, Fóruns de discussão da AWS e o AWS Support Center) e a documentação da AWS para instruir suas equipes. Entre em contato com o AWS Support por meio do AWS Support Center para obter ajuda com relação a dúvidas sobre a AWS. A AWS também compartilha as práticas recomendadas e os padrões que

aprendemos durante a operação da AWS na Amazon Builders' Library. Inúmeras outras informações úteis podem ser obtidas por meio do 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 em treinamento administrado por instrutor a fim de oferecer suporte adicional às suas equipes para o desenvolvimento de habilidades em serviços da 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. Provedores de serviços gerenciados como o AWS Managed Services e o AWS Managed Services Partners ou os provedores de serviços gerenciados na Rede de Parceiros da AWS fornecem especialização na implementação de ambientes de nuvem e atendem aos seus requisitos de segurança e conformidade e objetivos de negócios. 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 concentram-se nessas considerações de excelência operacional. (Para obter uma lista de perguntas e melhores práticas de excelência operacional, consulte 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 proprieda de, 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 atualize-as conforme as necessidades mudarem. Quando a responsabilidade 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-las para fornecer insights sobre seu status e criar os procedimentos para oferecer suporte a elas.

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 melhorem o fluxo de alterações na produção e permitam refatoração, comentários rápidos sobre a qualidade e correção de bugs. 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. Mantenha-se a par 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 workload (aplicações, infraestrutura, políticas, governança e operações) como 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, com níveis crescentes de controle de operações.

As perguntas a seguir concentram-se nessas considerações de excelência operacional.

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 recuperaç ão 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 sobre as métricas coletadas das workloads e nativamente na AWS. Você pode utilizar o CloudWatch ou aplicações de terceiros para agregar e apresentar visualizações das atividades operacionais em nível de negócios, workloads e operações. A AWS fornece insights sobre as workloads por meio de recursos de registro em log como o AWS X-Ray, o CloudWatch, o CloudTrail e o VPC Flow Logs, que possibilitam a identificação de problemas nas workloads, a fim de ajudar na análise e correção da causa raiz.

As perguntas a seguir concentram-se nessas considerações de excelência operacional.

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 de negócios 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 dados de log para o Amazon S3 ou enviar logs diretamente ao Amazon S3 para armazenamento de longo prazo. Usando o AWS Glue, você pode descobrir e preparar os dados de log no Amazon S3 para análise, e armazenar metadados associados no AWS Glue Data Catalog. Em seguida, você pode usar o Amazon Athena, por meio da integração nativa com o AWS Glue, para analisar os dados de log e consultá-los com o uso da linguagem SQL padrão. Usar 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.

A pergunta a seguir concentra-se nessas considerações de excelência operacional.

OPS 11: Como você faz para que as operações evoluam?

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 recursos a seguir para saber mais sobre as práticas recomendadas da AWS para Excelência operacional.

Documentação

- DevOps e AWS

Whitepaper

- Pilar Excelência operacional

Vídeo

- DevOps na Amazon

Segurança

O pilar Segurança refere-se à capacidade de proteger dados, sistemas e ativos para utilizar as tecnologias de nuvem para melhorar sua segurança.

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

Tópicos

- Princípios de design
- Definição
- Práticas recomendadas
- Recursos

Princípios de design

Existem sete princípios de design para segurança 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 por meio dos 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 no 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 seis áreas de práticas recomendadas de segurança na nuvem:

- Segurança
- Gerenciamento de identidade e acesso
- 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 sustenta nossos serviços de nuvem, você, como cliente da AWS, 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.

Práticas recomendadas

Tópicos

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.

A pergunta a seguir concentra-se nessas considerações sobre segurança. (Para obter uma lista de perguntas e melhores práticas de segurança, consulte 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 excelênci a operacional em nível de carga de trabalho e também organizacional e aplique-os a todas as áreas. Manter-se em dia com as recomendações da AWS, as fontes 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 workloads diferentes por conta, com base na respectiva função e nos requisitos de conformidade ou confidencialidade de dados, é uma abordagem recomendada.

Gerenciamento de identidade e acesso

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 oferecido principalmente pelo serviço AWS Identity and Access Management (IAM), que permite controlar o acesso programático e do usuário a serviços 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 workloads 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 sobre segurança.

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

Há dois tipos de identidade que você precisa gerenciar para operar workloads 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: seus administradores, desenvolvedores, operadores e usuários finais precisam de uma identidade para acessar seus ambientes e aplicações na AWS. Eles são membros de sua organização ou usuários externos com quem você colabora e que interagem com seus recursos da AWS por meio de um navegador da Web, de uma aplicação cliente ou de ferramentas interativas de linha de comando.

Identidades de máquina: suas aplicações de serviço, ferramentas operacionais e workloads precisam de uma identidade para fazer solicitações a serviços da AWS para ler dados, por exemplo. Essas identidades incluem máquinas em execução em seu ambiente da AWS, como instâncias do Amazon EC2 ou funções do AWS Lambda. Você também pode gerenciar identidad

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

es 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 workload. 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 de API a serviços da AWS, deve ser realizado usando credenciais de privilégio limitado e temporárias, como aquelas emitidas pelo AWS Security Token Service.

A AWS fornece recursos que podem ajudar você no gerenciamento de identidade e acesso. Para ajudá-lo a aprender melhores práticas, explore nossos laboratórios práticos sobre gerenciamento de credenciais e autenticação, controle de acesso humanoe 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 por meio do processamento de logs, eventos e monitoramentos que permitem auditoria, análises automatizadas e alarmes. Os logs do

CloudTrail, as chamadas de API da AWS e o CloudWatch fornecem monitoramento de métricas com alarmes e o AWS Config fornece um histórico de configuração. O Amazon GuardDuty é um serviço gerenciado de detecção de ameaças que monitora continuamente comportamentos mal-intencionados ou não autorizados para ajudar a proteger contas e workloads 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.

A pergunta a seguir concentra-se nessas considerações sobre segurança.

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 analisar os logs e responder a eles para que você possa identificar possíveis incidentes de segurança. A AWS fornece uma funcionalidade que torna o gerenciamento de logs mais fácil de implementar porque possibilita que você defina um ciclo de vida de retenção de dados ou em que local os dados serão preservados, 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 com estado e sem estado, seja usando tecnologias nativas da AWS ou produtos e serviços de parceiros disponíveis por meio do AWS Marketplace. Você deve usar a 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 sobre segurança.

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. 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 podem personalizar ou fortalecer a configuração de um Amazon Elastic Compute Cloud (Amazon EC2), de um contêiner do Amazon Elastic Container Service (Amazon ECS) ou de uma instância do AWS Elastic Beanstalk e persistir 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 facilita a criptografia e o gerenciamento de 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 oferecer um nível de 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 de gerenciamento de ciclo de vida de dados maior, 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 naquela Regiãor a menos que você explicitamente habilite um recurso ou utilize um serviço que forneça essa funcionalidade.

As perguntas a seguir se concentram nessas considerações sobre segurança.

SEC 7: Como você classifica seus 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 para 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 pelo 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. Colocar as ferramentas e o acesso 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:

- Está disponível o registro em log detalhado com conteúdo importante, como acesso e alterações a arquivo.
- 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.

A pergunta a seguir concentra-se nessas considerações sobre segurança.

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 de segurança.

Documentação

- Segurança na Nuvem AWS
- Conformidade da AWS
- Blog de segurança da AWS

Whitepaper

- Pilar Segurança
- Visão geral de segurança da AWS
- Risco e conformidade da AWS

Vídeo

- AWS Security State of the Union (Palestra sobre segurança da AWS)
- Visão geral de responsabilidade compartilhada

Confiabilidade

O pilar Confiabilidade abrange a capacidade de uma carga de trabalho de executar a função pretendida correta e consistentemente quando esperado. Isso inclui a capacidade de operar e testar a carga de trabalho durante todo o ciclo de vida dela. Este documento fornece orientações detalhadas sobre as práticas recomendadas para a implementação de workloads confiáveis na AWS.

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

Tópicos

- Princípios de design
- Definição

- Práticas recomendadas
- Recursos

Princípios de design

Existem cinco princípios de design para confiabilidade na nuvem:

- Recuperação automática de falhas: Ao monitorar indicadores-chave de performance (KPIs) 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.
- Testar os procedimentos de recuperação: em um ambiente on-premises, 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.
- Parar de tentar adivinhar a capacidade: uma causa comum de falha nas cargas de trabalho on-premises 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 Gerenciar cotas e restrições do Service Quotas.
- Gerencie as alterações na automação: 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 quatro áreas de melhores práticas para confiabilidade na nuvem:

- Fundamentos
- Arquitetura da carga de trabalho
- Gerenciamento de mudanças
- 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.

Práticas recomendadas

Tópicos

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 abordada 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 sobre confiabilidade. (Para obter uma lista de perguntas e melhores práticas de confiabilidade, consulte 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 infraestr utura de datacenter existente. Os planos devem incluir considerações de rede, como conectivi dade 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 as decisões iniciais de projeto que envolvem tanto o software quanto a infraestrutura. As decisões de arquitetura afetarão o comportamento da workload em todos os pilares do Well-Architected. Para atingir a confiabilidade, há padrões específicos que você deve seguir.

Com a AWS, os desenvolvedores de workload podem usar as linguagens e tecnologias que preferem. Os AWS SDKs 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 desenvolvedores também podem ler e descobrir como a Amazon cria e opera softwares na Amazon Builders' Library.

As perguntas a seguir se concentram nessas considerações sobre confiabilidade.

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 component es, 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.

Gerenciamento de alterações

As alterações na carga de trabalho ou no respectivo ambiente 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 as internas, como implantações de recursos e patches de segurança.

Com a AWS, você pode monitorar o comportamento de uma workload 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 sobre confiabilidade.

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 important es. 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 automatic amente 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 sobre confiabilidade.

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 Tempo médio até a recuperação (MTTR) 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. RTO e RPO são os seus objetivos para a restauração da workload. 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. A probabilidade de interrupção e o custo de recuperação também são fatores principais que ajudam a determinar o valor empresarial de fornecer a recuperação de desastres para uma workload.

Faça backup regular dos dados e teste os arquivos de backup para garantir a capacidade de 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, bem como objetivo de tempo de recuperação (RTO) e o objetivo de ponto de recuperação (RPO), para avaliar a resiliência de uma workload (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 de confiabilidade.

Documentação

- Documentação da AWS
- Infraestrutura global da AWS
- AWS Auto Scaling: como funcionam os planos de escalabilidade
- O que é o AWS Backup?

Whitepaper

- Pilar Confiabilidade: AWS Well-Architected
- Implementação de microsserviços na AWS

Eficiência de performance

O pilar Eficiência de performance inclui 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.

O pilar Eficiência de performance fornece uma visão geral dos princípios, melhores práticas e perguntas atinentes ao projeto. Você pode encontrar orientações prescritivas sobre implementação no Whitepaper sobre pilar de eficiência de performance.

Tópicos

- Princípios de design
- Definição
- Práticas recomendadas
- Recursos

Princípios de design

Existem cinco princípios de design para eficiência de performance na nuvem.

- Democratizar tecnologias avançadas: facilite a implementação de tecnologias avançadas 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 da workload em várias regiões da AWS ao redor do mundo permite que você forneça baixa latência e uma experiência melhor para os clientes por um custo mínimo.
- Usar arquiteturas sem servidor: 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 gerenciamento 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 afinidade mecânica: 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 quatro áreas de práticas recomendadas para eficiência de performance 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.

Se você analisar suas opções regularmente, terá certeza de que está se beneficiando da evolução contínua da Nuvem AWS. O monitoramento garante que você esteja ciente de qualquer desvio em relação à performance esperada. Faça concessões em sua arquitetura visando o aprimoramento da performance, como o uso de compactação ou armazenamento em cache, ou ainda a diminuição dos requisitos de consistência.

Práticas recomendadas

Tópicos

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.

A AWS disponibiliza recursos de vários tipos e em configurações diferentes. Desse modo, é mais fácil encontrar uma abordagem que atenda melhor às necessidades da workload. Você também pode encontrar opções que não são facilmente obtidas com infraestrutura no local. Por exemplo, um serviço gerenciado. como o Amazon DynamoDB, 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 sobre a eficiência da performance. (Para obter uma lista de perguntas e melhores práticas sobre eficiência de performance, consulte 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 ajudar você a selecionar uma arquitetura com base em conhecimentos do setor, mas os dados obtidos por meio de avaliações comparativas ou testes de carga serão necessários para otimizar a 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 específicos para a otimização da performance da 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 é disponibilizada em três formatos: instâncias, contêineres e funções:

- Instâncias são servidores virtualizados que permitem que você altere seus recursos com um simples botão ou uma única 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 unidades de estado sólido (SSDs) e unidades de processamento gráfico (GPUs).
- 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).

- 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 sobre a eficiência da performance.

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 é disponibilizado em três formatos: objeto, bloco e arquivo:

- Armazenamento de objeto 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 9s) de durabilidade e armazena dados para milhões de aplicativos de empresas em todo o mundo.

- O Armazenamento em bloco fornece 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 SAN. O Amazon Elastic Block Store (Amazon EBS) foi projetado para cargas de trabalho que exigem armazenamento persistente acessível por instâncias do EC2 e que ajuda você a ajustar aplicativos com os níveis ideais de capacidade de armazenamento, performance e custo.
- 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), 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 mais simples e econômico o processo de execução de sistemas de arquivos conhecidos, para que você possa aproveitar os conjuntos de recursos avançados e a rápida performance de sistemas de arquivos de código aberto amplamente utilizados e licenciados comercialmente.

As perguntas a seguir se concentram nessas considerações sobre a eficiência da performance.

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 rendiment o necessário, a frequência de acesso (online, offline, arquivamento), a frequência de atualizaç ão (WORM, dinâmica) e as restrições de disponibilidade e durabilidade. Os sistemas Well-Arch itected 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 com propósito específico, inclusive bancos de dados relacionais, ledger, de chave-valor, de documentos, de grafos, de séries temporais e em memória. Com os bancos de dados da AWS, você não precisa se preocupar com tarefas de gerenciamento de banco de dados, como provisionamento de servidor, aplicação de patches, instalação, configuração, backups ou recuperação. A AWS monitora continuamente os clusters para manter as workloads em pleno funcionamento por meio de armazenamento com recuperação automática e escalabilidade automatizada, para que você se concentre no desenvolvimento de aplicações de valor superior.

As perguntas a seguir se concentram nessas considerações sobre a eficiência da performance.

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. A seleção da solução e dos recursos de banco de dados incorretos para um sistema pode levar a uma menor performance do sistema.

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).

da AWS

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. Desse modo, fica mais fácil compatibilizar seus métodos de rede com suas necessidades. A AWS oferece recursos de produtos (por exemplo, redes avançadas, instâncias otimizadas do Amazon EBS, aceleração de transferências do Amazon S3 e a dinâmica do Amazon CloudFront) para otimizar o tráfego da rede. A AWS também oferece recursos de rede (por exemplo, roteamento de latência do Amazon Route 53, endpoints da Amazon VPC, AWS Direct Connect e AWS Global Accelerator) para reduzir a distância ou a oscilação da rede.

As perguntas a seguir se concentram nessas considerações sobre a eficiência da performance.

PERF 5: Como você configura sua solução de rede?

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. É possível optar por colocar os recursos perto de onde eles serão usados para reduzir a distância. Use métricas de rede para fazer alterações na configuração de rede conforme a carga de trabalho evolui. Ao aproveitar as Regiões, grupos de posicionamento e serviços de borda, você pode melhorar a performance significativamente. 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 estão evoluindo rapidamente, e você deve garantir que os componentes da carga de trabalho estejam usando as tecnologias e abordagens mais recentes 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 (ML) e inteligência artificial (IA), podem permitir a reconstrução das experiências do cliente e inovações em todas as cargas de trabalho de negócios.

Aproveite a inovação contínua na AWS, orientada 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 sobre a eficiência da performance.

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.

Em geral, 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 observabilidade que fornece dados e insights práticos para monitorar workloads, responder a alterações de performance em todo o sistema, otimizar a utilização de recursos e obter uma visão unificada da integridade operacional. O CloudWatch coleta dados operacionais e de monitoramento em forma de logs, métricas e eventos com base em workloads executadas na AWS e em servidores on-premises. O AWS X-Ray ajuda os desenvolvedores a analisar e depurar aplicações distribuídas em produção. Com o AWS X-Ray, você pode obter insights sobre a performance do seu 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 sobre a eficiência da performance.

PERF 7: Como você monitora seus recursos para garantir que eles estejam apresentando boa performance?

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 ficar mais próximo dos 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 sobre a eficiência da performance.

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 eficiência de performance.

Documentação

- Otimização da performance do Amazon S3
- Performance de volumes do Amazon EBS

Whitepaper

- Pilar Eficiência de performance

Vídeo

- AWS re:Invent 2019: Amazon EC2 foundations (Fundamentos do Amazon EC2) (CMP211-R2)
- AWS re:Invent 2019: Leadership session: Storage state of the union (Sessão de liderança: palestra sobre armazenamento) (STG201-L)
- AWS re:Invent 2019: Leadership session: AWS purpose-built databases (Sessão de liderança: bancos de dados com propósito específico) (DAT209-L)
- AWS re:Invent 2019: Connectivity to AWS and hybrid AWS network architectures (Conectividade com a AWS e arquiteturas de rede híbridas da AWS) (NET317-R1)
- AWS re:Invent 2019: Powering next-gen Amazon EC2: Deep dive into the Nitro system (Potencialização de última geração: aprofundamento sobre o sistema Nitro) (CMP303-R2)
- AWS re:Invent 2019: Scaling up to your first 10 million users (Aumente a escala verticalmente para atingir seus primeiros dez milhões de usuários) (ARC211-R)

Otimização de custos

O pilar Otimização de custos inclui a capacidade de executar sistemas para proporcionar valor comercial pelo menor preço.

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 sobre o pilar de otimização de custos.

Tópicos

Princípios de design

Existem cinco princípios de design para otimização de custos na nuvem:

- Implemente 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 Excelência operacional, você precisa criar aptidão por meio da criação de conhecimento, programas, recursos e processos para se tornar uma organização econômica.
- Adote um modelo de consumo: pague somente pelos recursos de computação necessários e aumente ou reduza o uso dependendo dos requisitos comerciais, sem usar 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).
- Avalie 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 realiza as tarefas pesadas que não geram diferenciação das operações de datacenter, como armazenamento em rack, empilhamento e 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.
- Analise e atribua 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 cinco áreas de práticas recomendadas para otimização de custos na nuvem:

- Pratique 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 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.

Práticas recomendadas

Tópicos

Pratique 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 AWS Budgets fornece notificações proativas para custo e uso. Os Blogs da AWS oferecem informações sobre novos serviços e recursos para garantir que você se mantenha em dia com os novos lançamentos de serviços.

As perguntas a seguir concentram-se nessas considerações sobre otimização de custos. (Para obter uma lista de perguntas e melhores práticas de otimização de custos, consulte o Apêndice.).

COST 1: Como implementar o gerenciamento financeiro na nuvem?

A implementação do gerenciamento financeiro na nuvem possibilita que as organizações obtenham valor empresarial e sucesso financeiro à medida que elas otimizam os custos e o uso e escalam na AWS.

Ao criar uma função de otimização de custos, use membros e complemente a equipe com especialistas em CFM e otimização de custos. Os membros existentes da equipe compreenderão 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 na sua organização, melhore ou desenvolva programas e processos existentes. É muito mais rápido adicionar ao que já existe do que criar novos 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 contas com o AWS Organizations ou o AWS Control Tower, o que fornece separação de contas e ajuda na alocação de custos e uso. Você também pode usar a marcação de 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 de custos e de uso é feito com notificações, por meio do AWS Budgets, e de controles, por meio do AWS Identity and Access Management (IAM) e do Service Quotas.

As perguntas a seguir concentram-se nessas considerações sobre otimização de custos.

COST 2: Como você controla 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 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 etiquetas de alocação de custos para categorizar e monitorar o uso e os custos da AWS. Quando você aplica etiquetas aos recursos da AWS (como instâncias do EC2 ou buckets do S3), a AWS gera um relatório de custos e uso com base em suas etiquetas e utilização. 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 análises e inspeções mais profundas, use a granularidade por hora no AWS Cost Explorer ou o Amazon Athena e o Amazon QuickSight com o Relatório de Custos 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 Amazon EC2 oferece uma variedade de opções de preço flexíveis e econômicas para você adquirir instâncias do AWS e de outros serviços que sejam mais adequados às suas necessidades. Sob demanda Instâncias permitem que você pague pela capacidade de computação por hora, sem nenhum requisito mínimo de comprometimento. Savings Plans e Instâncias reservadas oferecem economias de até 75% da 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. 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, como ao usar o Amazon Aurora em RDS para remover gastos com licenças caras de banco de dados.

As perguntas a seguir concentram-se nessas considerações sobre otimização de custos.

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

O Amazon EC2, o Amazon EBS e o Amazon S3 são serviços fundamentais da AWS. Serviços gerenciados como o Amazon RDS e o Amazon DynamoDB são serviços da AWS de nível superior ou em nível de aplicação. 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 administr ativa 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 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 o Cost Explorer e o 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 os recursos automaticamente para que correspondam à demanda da workload. 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 utilização ou o Amazon SQS para implementar uma fila na sua carga de trabalho. Os dois permitirão que você modifique a demanda nos componentes da carga de trabalho.

As perguntas a seguir concentram-se nessas considerações sobre otimização de custos.

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 utilizaçã o distorcida em ambas as direções tem um impacto adverso sobre a organização, tanto nos custos operacionais (redução na performance em decorrência de utilização excessiva) quanto em despesas desnecessárias na 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

À medida que a AWS lança novos serviços e recursos, uma das práticas recomendadas é avaliar suas decisões sobre a arquitetura existente a fim de garantir que elas ofereçam o melhor custo-benefício. 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 concentram-se nessas considerações sobre otimização de custos.

COST 10: Como você avalia os novos serviços?

À medida que a AWS lança novos serviços e recursos, uma das práticas recomendadas é avaliar suas decisões sobre a arquitetura existente a fim de garantir que elas ofereçam o melhor custo-benefício.

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 relacionais. O uso de recursos sem servidor, como o Lambda, pode remover a necessidade de operar e gerenciar instâncias para executar código.

Recursos

Consulte os recursos a seguir para saber mais sobre nossas melhores práticas de otimização de custos.

Documentação

- Documentação da AWS

Whitepaper

- Pilar Otimização de custos

Sustentabilidade

O pilar Sustentabilidade focaliza os impactos ambientais, especialmente a eficiência e o consumo de energia, que são fatores importantes para fundamentar ações diretas dos arquitetos destinadas a reduzir o uso de recursos. Você pode encontrar orientações prescritivas sobre implementação no whitepaper sobre o pilar de sustentabilidade.

Tópicos

- Princípios de design
- Definição
- Práticas recomendadas

Princípios de design

Existem seis princípios de design para sustentabilidade na nuvem.

- Compreenda seu impacto: Meça o impacto da seu workload na nuvem e modele seu impacto futuro. Inclua todas as fontes de impacto, inclusive aquelas resultantes do uso de seus produtos pelo cliente e da desativação e descontinuação deles. Compare o resultado produtivo com o

impacto total de suas workloads em nuvem analisando os recursos e as emissões exigidas por unidade de trabalho. Use esses dados para estabelecer indicadores-chave de performance (KPIs), avaliar maneiras de melhorar a produtividade enquanto reduz o impacto e estimar o impacto das mudanças propostas ao longo do tempo.

- Estabeleça metas de sustentabilidade: Para cada workload em nuvem, estabeleça metas de sustentabilidade de longo prazo, tais como reduzir os recursos de computação e armazenamento exigidos por transação. Modele o retorno sobre o investimento para as melhorias de sustentabilidade das workloads e ofereça aos proprietários os recursos de que eles precisam para investir em metas de sustentabilidade. Planeje-se para o crescimento e projete suas workloads de forma que seu desenvolvimento resulte em uma intensidade de impacto menor com relação a uma unidade apropriada, como por usuário ou por transação. As metas ajudam você a respaldar os objetivos de sustentabilidade mais amplos de sua empresa ou organização, identificar regressões e priorizar áreas para possível melhoria.
- Maximize a utilização: Dimensione as workloads corretamente e implemente um design eficiente que garanta uma alta utilização e maximize a eficiência de energia do hardware subjacente. Dois hosts com 30% de utilização são menos eficientes do que um host com 60% devido ao consumo de energia de referência por host. Ao mesmo tempo, elimine ou minimize recursos, processamento e armazenamento ociosos para reduzir a energia total necessária para suprir a workload.
- Antecipe e adote ofertas de hardware e software novos e mais eficientes: Apoie as melhorias preventivas que seus parceiros e fornecedores fazem para ajudar você a reduzir o impacto das workloads em nuvem. Monitore e avalie continuamente as ofertas de software e hardware novos e mais eficientes. Projete visando a flexibilidade para permitir a adoção rápida de novas tecnologias eficientes.
- Use serviços gerenciados: Compartilhar serviços com uma ampla base de clientes ajuda a maximizar a utilização de recursos, o que reduz a quantidade de infraestrutura necessária para comportar as workloads em nuvem. Por exemplo, os clientes podem compartilhar o impacto de componentes comuns de um datacenter, como energia e redes, migrando workloads para a Nuvem AWS e adotando serviços gerenciados como o AWS Fargate para contêineres com tecnologia sem servidor, os quais são operados em escala pela AWS, que é responsável pela eficiência da operação. Use serviços gerenciados que possam ajudar a minimizar seu impacto, como a migração automática de dados acessados com pouca frequência para o armazenamento com pouco acesso com as configurações do ciclo de vida do Amazon S3 ou o Amazon EC2 Auto Scaling para ajustar a capacidade de acordo com a demanda.
- Reduza o impacto posterior de suas workloads na nuvem Reduza a quantidade de energia ou recursos necessários para usar seus serviços. Reduza ou elimine a necessidade de os clientes

fazerem upgrade de dispositivos para usar seus serviços. Teste o uso de farms de dispositivos para saber qual é o impacto esperado e teste com os clientes para entender o impacto atual do uso de seus serviços.

Definição

Existem seis áreas de práticas recomendadas de sustentabilidade na nuvem.

- Escolha de região
- Padrões de comportamento do usuário
- Padrões de software e arquitetura
- Padrões de dados
- Padrões de hardware
- Processo de desenvolvimento e implantação

A sustentabilidade na nuvem é uma iniciativa contínua direcionada principalmente à redução do consumo de energia e à eficiência energética em todos os componentes de uma workload por meio da maximização dos benefícios dos recursos provisionados e da minimização do total de recursos necessários. Essa iniciativa pode incluir vários fatores, como seleção inicial de uma linguagem de programação eficiente, adoção de algoritmos modernos, uso de técnicas eficientes de armazenamento de dados, implantação em uma infraestrutura de computação eficiente e corretamente dimensionada e minimização dos requisitos de hardware do usuário final com alto consumo de energia.

Práticas recomendadas

Tópicos

Escolha de região

Escolha as regiões onde você vai implementar suas workloads com base em seus requisitos empresariais e em suas metas de sustentabilidade.

A pergunta a seguir concentra-se nessas considerações sobre sustentabilidade. (Para obter uma lista de perguntas e práticas recomendadas de sustentabilidade, consulte o Apêndice.)

SUS 1: Como você escolhe as regiões para apoiar suas metas de sustentabilidade?

Escolha regiões próximas aos projetos de energia renovável da Amazon e regiões onde a grade de intensidade de carbono publicada esteja abaixo de outros locais (ou regiões).

Padrões de comportamento do usuário

A maneira como os usuários consomem suas workloads e outros recursos pode ajudar você a identificar melhorias para atingir as metas de sustentabilidade. Escale a infraestrutura de tal forma que ela sempre corresponda à carga de usuários e implante apenas os recursos mínimos necessários para oferecer suporte aos usuários. Alinhe os níveis de serviço às necessidades do cliente. Posicione os recursos de maneira a limitar a rede necessária para que eles sejam consumidos pelos usuários. Remova ativos que não sejam utilizados. Identifique ativos criados que não são utilizados e pare de gerá-los. Forneça aos membros da sua equipe dispositivos compatíveis com suas necessidades e com impacto de sustentabilidade reduzido.

A pergunta a seguir concentra-se nessas considerações sobre sustentabilidade:

SUS 2: Como você aproveita os padrões de comportamento do usuário para apoiar suas metas de sustentabilidade?

A maneira como os usuários consomem suas workloads e outros recursos pode ajudar você a identificar melhorias para atingir as metas de sustentabilidade. Escale a infraestrutura de tal forma que ela sempre corresponda à carga de usuários e implante apenas os recursos mínimos necessários para oferecer suporte aos usuários. Alinhe os níveis de serviço às necessidades do cliente. Posicione os recursos de maneira a limitar a rede necessária para que eles sejam consumidos pelos usuários. Remova ativos que não sejam utilizados. Identifique ativos criados que não são utilizados e pare de gerá-los. Forneça aos membros da sua equipe dispositivos compatíveis com suas necessidades e com impacto de sustentabilidade reduzido.

Escale a infraestrutura com a carga de usuários: identifique períodos de baixa utilização ou em que não há utilização e escale os recursos para eliminar a capacidade em excesso e melhorar a eficiência.

Alinhar SLAs com os objetivos de sustentabilidade: defina e atualize as metas dos Acordos de Serviço (SLAs), como períodos de disponibilidade ou de retenção de dados a fim de minimizar o número de recursos exigidos para comportar as workloads e, ao mesmo tempo, continuar atendendo aos requisitos empresariais.

Elimine a criação e a manutenção de ativos ociosos: analise os ativos de aplicações (como relatórios pré-compilados, conjuntos de dados e imagens estáticas) e os padrões de acesso aos ativos para identificar redundâncias, subutilização e possíveis alvos de desativação. Consolide ativos gerados com conteúdo redundante (por exemplo, relatórios mensais com saídas e conjuntos de dados que se sobreponham ou sejam comuns) para eliminar os recursos consumidos quando há duplicação de saídas. Desative ativos não utilizados (por exemplo, imagens de produtos que não são mais vendidos) para liberar os recursos consumidos e reduzir o número de recursos usados para comportar a workload.

Otimize o posicionamento geográfico das workloads de acordo a localização dos usuários: analise os padrões de acesso à rede para identificar de onde seus clientes estão se conectando geograficamente. Escolha regiões e serviços que reduzam a distância que o tráfego de rede deve percorrer para reduzir o total de recursos de rede necessários para comportar a workload.

Otimize os recursos dos membros da equipe para as atividades executadas: otimize os recursos fornecidos aos membros da equipe para minimizar o impacto sobre a sustentabilidade e, ao mesmo tempo, atender às necessidades deles. Por exemplo, realize operações complexas, como renderização e compilação, em desktops compartilhados na nuvem com alta utilização em vez de em sistemas de usuário único subutilizados com alto consumo de energia.

Padrões de software e arquitetura

Implemente padrões que suavizem os picos de carga e mantenham a alta utilização consistente de recursos implantados para minimizar os recursos consumidos. Os componentes podem ficar ociosos devido à falta de uso por conta das mudanças no comportamento do usuário ao longo do tempo. Revise os padrões e a arquitetura para consolidar os componentes subutilizados a fim de aumentar a utilização geral. Retire os componentes que não são mais necessários. Saiba qual é a performance dos componentes de sua workload e otimize os componentes que consomem a maioria dos recursos. Esteja ciente dos dispositivos que seus clientes usam para acessar seus serviços e implemente padrões a fim de minimizar a necessidade de upgrades de dispositivos.

As perguntas a seguir se concentram nessas considerações sobre sustentabilidade:

SUS 3: Como você aproveita os padrões de software e arquitetura para apoiar suas metas de sustentabilidade?

Implemente padrões que suavizem os picos de carga e mantenham a alta utilização consistente de recursos implantados para minimizar os recursos consumidos. Os componentes podem ficar ociosos devido à falta de uso por conta das mudanças no comportamento do usuário ao longo do tempo. Revise os padrões e a arquitetura para consolidar os componentes subutilizados a fim de aumentar a utilização geral. Retire os componentes que não são mais necessários. Saiba qual é a performance dos componentes de sua workload e otimize os componentes que consomem a maioria dos recursos. Esteja ciente dos dispositivos que seus clientes usam para acessar seus serviços e implemente padrões a fim de minimizar a necessidade de upgrades de dispositivos.

Otimize o software e a arquitetura para trabalhos assíncronos e programados: use designs e arquiteturas eficientes de software para minimizar a média de recursos necessários por unidade de trabalho. Implemente mecanismos que resultem em uma utilização uniforme de componentes para reduzir os recursos ociosos entre as tarefas e minimizar o impacto de picos de carga.

Remova ou refatore os componentes da workload com baixa utilização ou que não estão sendo usados: monitore a atividade da workload para identificar alterações na utilização de componentes individuais ao longo do tempo. Remova os componentes que não são mais utilizados nem necessários e refatore os componentes pouco usados para reduzir o desperdício de recursos.

Otimize as áreas de código que mais consomem tempo e recursos: monitore a atividade da workload para identificar os componentes da aplicação que mais consomem recursos. Otimize o código que é executado nesses componentes para minimizar o uso de recursos e, ao mesmo tempo, maximizar a performance.

Otimize o impacto sobre os dispositivos e o equipamento do cliente: conheça os dispositivos e o equipamento que os clientes usam para consumir seus serviços, o ciclo de vida esperado para eles e o impacto financeiro e na sustentabilidade decorrente da substituição desses componentes. Implemente padrões e arquiteturas de software de modo a minimizar a necessidade de substituir dispositivos e fazer upgrade de equipamento. Por exemplo, implemente novos recursos usando código compatível com versões anteriores de sistemas operacionais e hardware mais antigos ou gerencie o tamanho das cargas úteis para que elas não excedam a capacidade de armazenamento do dispositivo de destino.

Use padrões de software e arquiteturas que comportem melhor os padrões de acesso a dados e de armazenamento: entenda como os dados são usados dentro da workload, consumidos pelos usuários, transferidos e armazenados. Escolha tecnologias com o mínimo de requisitos de armazenamento e processamento de dados.

Padrões de dados

Implemente padrões que suavizem os picos de carga e mantenham a alta utilização consistente de recursos implantados para minimizar os recursos consumidos. Os componentes podem ficar ociosos devido à falta de uso por conta das mudanças no comportamento do usuário ao longo do tempo. Revise os padrões e a arquitetura para consolidar os componentes subutilizados a fim de aumentar a utilização geral. Retire os componentes que não são mais necessários. Saiba qual é a performance dos componentes de sua workload e otimize os componentes que consomem a maioria dos recursos. Esteja ciente dos dispositivos que seus clientes usam para acessar seus serviços e implemente padrões a fim de minimizar a necessidade de upgrades de dispositivos.

A pergunta a seguir concentra-se nessas considerações sobre sustentabilidade:

SUS 4: Como você aproveita o acesso a dados e os padrões de uso para apoiar suas metas de sustentabilidade?

Implemente práticas de gerenciamento de dados para reduzir o armazenamento provisionado necessário para comportar a workload e os recursos exigidos para usá-la. Entenda seus dados e use as tecnologias e as configurações de armazenamento que melhor promovam o valor empresarial dos dados e a forma como eles são usados. Gerencie o ciclo de vida dos dados e opte por um armazenamento mais eficiente e com menor performance quando os requisitos diminuírem, excluindo os dados que não são mais necessários.

Implemente uma política de classificação de dados: classifique os dados para entender o significado deles para os resultados dos negócios. Use essas informações para determinar quando é possível migrar os dados para um armazenamento com uso mais eficiente de energia ou excluí-los de forma segura.

Use tecnologias que comportem os padrões de acesso a dados e armazenamento: use um armazenamento mais adequado à maneira como os dados são acessados e armazenados a fim de reduzir os recursos provisionados e, ao mesmo tempo, atender à sua workload. Por exemplo, dispositivos de estado sólido (SSDs) usam mais energia do que unidades magnéticas e devem

ser usados somente para casos de uso de dados ativos. Use um armazenamento de classe de arquivamento com eficiência de energia para dados acessados com pouca frequência.

Use políticas de ciclo de vida para excluir dados desnecessários: gerencie o ciclo de vida de todos os dados e defina cronogramas de exclusão automática para minimizar os requisitos totais de armazenamento da workload.

Minimize o provisionado em excesso no armazenamento em bloco: para reduzir o armazenamento total provisionado, crie um armazenamento em bloco com alocações por tamanho que sejam apropriadas à workload. Use volumes elásticos para expandir o armazenamento à medida que os dados aumentam sem precisar redimensionar o armazenamento anexado aos recursos de computação. Analise regularmente volumes elásticos e reduza volumes com excesso de provisionamento para se ajustar ao tamanho de dados atual.

Remova dados desnecessários ou redundantes: duplique os dados somente quando necessário para reduzir o armazenamento total consumido. Use tecnologias de backup que eliminem dados duplicados em níveis de arquivo e bloco. Limite o uso de configurações RAID (Matriz redundante de unidades independentes), exceto quando necessário para atender aos SLAs.

Use sistemas de arquivos compartilhados para acessar dados comuns: adote o armazenamento compartilhado e fontes únicas de verdade para evitar duplicação de dados e reduzir os requisitos totais de armazenamento da workload. Busque dados do armazenamento compartilhado somente conforme necessário. Desvincule volumes não usados para liberar recursos. Minimize a movimentação de dados entre redes: use o armazenamento compartilhado e acesse dados de datastores regionais para minimizar os recursos totais de rede exigidos para comportar a movimentação de dados da workload.

Faça backup dos dados somente quando for difícil recriar: para reduzir o consumo de armazenamento, faça backup somente de dados com valor empresarial ou que sejam necessários para atender aos requisitos de conformidade. Examine as políticas de backup e exclua armazenamentos temporários que não forneçam valor em um cenário de recuperação.

Padrões de hardware

Procure oportunidades para reduzir os impactos na sustentabilidade da workload fazendo mudanças nas suas práticas de gerenciamento de hardware. Minimize a quantidade de hardware necessária para provisionar e implantar e escolha o hardware mais eficiente para sua workload individual.

A pergunta a seguir concentra-se nessas considerações sobre sustentabilidade:

SUS 5: Como suas práticas de gerenciamento de hardware e de uso apoiam suas metas de sustentabilidade?

Procure oportunidades para reduzir os impactos na sustentabilidade da workload fazendo mudanças nas suas práticas de gerenciamento de hardware. Minimize a quantidade de hardware necessária para provisionar e implantar e escolha o hardware mais eficiente para sua workload individual.

Use uma quantidade mínima de hardware para atender às suas necessidades: usando os recursos da nuvem, é possível fazer alterações frequentes nas implementações da workload. Atualize os componentes implantados conforme suas necessidades mudarem.

Use tipos de instância cujo impacto seja mínimo: monitore continuamente o lançamento de novos tipos de instância e aproveite as melhorias de eficiência energética, incluindo os tipos de instância projetados para comportar workloads específicas, como treinamento e inferência de machine learning e transcodificação de vídeo.

Use serviços gerenciados: os serviços gerenciados transferem para a AWS a responsabilidade pela manutenção de uma média elevada de utilização e pela otimização da sustentabilidade do hardware implantado. Use serviços gerenciados para distribuir o impacto na sustentabilidade do serviço entre todos os locatários dele, reduzindo sua contribuição individual.

Otimize o uso de GPUs: as unidades de processamento gráfico (GPUs) podem ser uma fonte de alto consumo de energia e várias workloads de GPU são altamente variáveis, como renderização, transcodificação e treinamento e modelagem de machine learning. Execute instâncias de GPUs somente pelo tempo necessário e desative-as com automação quando não precisar mais delas para reduzir o consumo de recursos.

Padrões de desenvolvimento e implantação

Procure oportunidades para reduzir seu impacto na sustentabilidade fazendo mudanças nas suas práticas de desenvolvimento, teste e implantação.

A pergunta a seguir concentra-se nessas considerações sobre sustentabilidade:

SUS 6: Como seus processos de desenvolvimento e implantação apoiam suas metas de sustentabilidade?

Procure oportunidades para reduzir seu impacto na sustentabilidade fazendo mudanças nas suas práticas de desenvolvimento, teste e implantação.

Adote métodos que possam introduzir melhorias de sustentabilidade rapidamente: teste e valide possíveis melhorias antes de implantá-las na produção. Considere o custo do teste ao calcular o benefício futuro potencial de uma melhoria. Desenvolva métodos de teste de baixo custo para permitir pequenas melhorias.

Mantenha a workload atualizada: bibliotecas, aplicações e sistemas operacionais atualizados podem melhorar a eficiência da workload e facilitar a adoção de tecnologias mais eficientes. Um software atualizado também pode incluir recursos para medir o impacto na sustentabilidade da workload com mais precisão, pois os fornecedores oferecem recursos para atender às suas próprias metas de sustentabilidade.

Aumente a utilização dos ambientes de compilação: use automação e infraestrutura como código para ativar ambientes de pré-produção, quando necessário, e desativá-los quando não estiverem sendo usados. Um padrão comum é programar períodos de disponibilidade que coincidam com as horas de trabalho dos membros da equipe de desenvolvimento. A hibernação é uma ferramenta útil para preservar o estado e colocar rapidamente as instâncias online apenas quando necessário. Use tipos de instância com capacidade de intermitência, instâncias Spot, serviços de banco de dados elásticos, contêineres e outras tecnologias para alinhar a capacidade de desenvolvimento e teste com o uso.

Use farms de dispositivos gerenciados para testes: farms de dispositivos gerenciados distribuem o impacto na sustentabilidade da fabricação de hardware e do uso de recursos entre vários locatários. Farms de dispositivos gerenciados oferecem diversos tipos de dispositivos para que você ofereça compatibilidade com componentes de hardware mais antigos e menos populares e evite o impacto sobre a sustentabilidade do cliente devido a atualizações desnecessárias de dispositivos.

Recursos

Consulte os recursos a seguir para saber mais sobre nossas práticas recomendadas de sustentabilidade.

Whitepaper

- Pilar da sustentabilidade

Vídeo

- The Climate Pledge

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, e 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 Framework 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 análises devem ser aplicadas nos principais marcos do ciclo de vida do produto, logo no início da fase de projeto, para evitar portas de mão única difíceis de se alterar e antes da data de lançamento. (Muitas decisões são bidirecionais e reversíveis. Elas podem ser tomadas com um processo leve. As vias de mão única são difíceis ou impossíveis de reverter e requerem maior inspeção antes de serem feitas.) Depois que você entrar em produção, sua carga de trabalho continuará evoluindo, à medida que você adiciona novos recursos e altera implementações de tecnologias. 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 que todas as pessoas certas participem da conversa. Muitas vezes, descobrimos que as análises constituem 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 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 abordar os 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 uma final de campeonato, 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 manuais estratégicos para lidar com eles, caso ocorram.
- “Não queremos que outras pessoas saibam os segredos da implementação da nossa solução!”
  - Se você apresentar as perguntas do Well-Architected Framework aos membros da equipe, eles verão que nenhuma delas revela informações proprietárias comerciais ou técnicas.

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 grupos 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 oferece práticas recomendadas de arquitetura nos seis pilares para projetar e operar sistemas confiáveis, seguros, eficientes, econômicos e sustentáveis na nuvem. O Framework fornece um conjunto de perguntas que permite analisar uma arquitetura existente ou proposta. Ele também fornece um conjunto de práticas recomendadas 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

Os indivíduos e empresas a seguir contribuíram para este documento:

- Brian Carlson, líder de operações do Well-Architected, Amazon Web Services
- Ben Potter, Líder de Segurança do Amazon Web Services (AWS) Well-Architected
- Seth Eliot: líder de confiabilidade do Well-Architected, Amazon Web Services
- Eric Pullen arquiteto de soluções sênior, Amazon Web Services
- Rodney Lester, arquiteto-chefe de soluções, Amazon Web Services
- Jon Steele, Gerente técnico sênior de contas, Amazon Web Services
- Max Ramsay: arquiteto-chefe de soluções de segurança, Amazon Web Services
- Callum Hughes, arquiteto de soluções, Amazon Web Services
- Aden Leirer, gerente de programa de conteúdo do Well-Architected, Amazon Web Services

Leitura adicional

Centro de Arquitetura da AWS

Conformidade com a Nuvem AWS

Programa de parceiros do AWS Well-Architected

AWS Well-Architected Tool

Página inicial do AWS Well-Architected

whitepaper sobre o pilar de excelência operacional

whitepaper Pilar de segurança

whitepaper sobre o pilar de confiabilidade

Whitepaper sobre pilar de eficiência de performance

whitepaper sobre o pilar de otimização de custos

whitepaper sobre o pilar de sustentabilidade

Amazon Builders' Library

Revisões do documento

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

Alteração	Descrição	Data
Atualização secundária	Adição da definição de nível de esforço e atualização das práticas recomendadas no apêndice.	October 20, 2022
Whitepaper atualizado	Adição do pilar Sustentab ilidade e atualização dos links.	December 2, 2021
Atualização principal	Adição do pilar Sustentab ilidade ao Framework.	November 20, 2021
Atualização secundária	Eliminação de linguagem não inclusiva	April 22, 2021
Atualização secundária	Correção de vários links.	March 10, 2021
Atualização secundária	Pequenas alterações editoriai s.	July 15, 2020
Atualizações para a nova estrutura de trabalho	Revisão e reescrita da maioria das perguntas e respostas.	July 8, 2020
Whitepaper atualizado	Adição do AWS Well-Arch itected Tool, de links para os laboratórios do AWS Well-Architected e de parceiros do AWS Well-Architected, e correções secundárias para possibilitar uma versão em vários idiomas do Framework.	July 1, 2019
Whitepaper atualizado	Revisão e reescrita da maioria das perguntas e respostas,	November 1, 2018

	para garantir que as perguntas se concentrem em um tópico de cada vez. Isso fez com que algumas perguntas anteriore s 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.	
Whitepaper atualizado	Atualizações para simplificar o texto de pergunta, padroniza r respostas e melhorar a legibilidade.	June 1, 2018
Whitepaper atualizado	O trecho sobre excelência operacional foi movido para a frente dos pilares e reescrito para enquadrar outros pilares. Outros pilares foram atualizad os para refletir a evolução da AWS.	November 1, 2017
Whitepaper atualizado	Atualização do Framework para incluir o pilar de excelênci a operacional e revisão e atualização dos outros pilares para reduzir a duplicação e incorporar aprendizados da realização de análises com milhares de clientes.	November 1, 2016
Atualizações secundárias	Atualização do Apêndice com informações atuais do Amazon CloudWatch Logs.	November 1, 2015

Publicação inicial	Publicação do AWS Well-Architected Framework.	October 1, 2015

Apêndice: Perguntas e práticas recomendadas

Tópicos

Excelência operacional

Tópicos

Organização

Perguntas

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.

Práticas recomendadas

- OPS01-BP02 Avalie as necessidades dos clientes internos
- OPS01-BP03 Avaliar os requisitos de governança
- OPS01-BP04 Avaliar os requisitos de conformidade
- OPS01-BP05 Avaliar o cenário de ameaças
- OPS01-BP06 Avalie as compensações
- OPS01-BP07 Gerenciar os benefícios e os riscos

OPS01-BP01 Avaliar as necessidades dos 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.

Antipadrões comuns:

- Você decidiu não ter suporte ao cliente fora do horário comercial principal, mas não analisou dados históricos de solicitação de suporte. Você não sabe se isso afetará seus clientes.
- Você está desenvolvendo um novo recurso, mas não envolveu seus clientes para descobrir se ele é desejado, em qual formato é desejado e sem experimentação para validar a necessidade e o método de entrega.

Benefícios do estabelecimento desta prática recomendada: Os clientes cujas necessidades estão atendidas têm muito mais probabilidade de permanecerem como clientes. Avaliar e compreender as necessidades de clientes externos informará como você priorizará seus esforços para entregar valor empresarial.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Compreender as necessidades empresariais: o sucesso nos negócios é possibilitado pelos objetivos e pelo entendimento compartilhados entre as partes interessadas, incluindo equipes corporativas, de desenvolvimento e de operações.
  - Analisar os objetivos, as necessidades e as prioridades empresariais dos clientes externos: envolva as principais partes interessadas, incluindo as equipes corporativas, de

desenvolvimento e de operações, para discutir as metas, as necessidades e as prioridades dos clientes externos. Isso garantirá que você tenha um entendimento completo do suporte às operações que é necessário para obter resultados nos negócios.

- Estabelecer uma compreensão compartilhada: estabeleça uma compreensão compartilhada das funções corporativas sobre a workload, as funções de cada uma das equipes na operação da workload e de como esses fatores oferecem apoiam seus objetivos empresariais compartilhados entre os clientes internos e externos.

Recursos

Documentos relacionados:

- AWS Well-Architected Framework Concepts – Feedback loop (Conceitos do AWS Well-Architected Framework: loop de feedback)

OPS01-BP02 Avalie as necessidades dos clientes internos

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.

Use suas prioridades estabelecidas para concentrar seus esforços de melhoria onde eles terão maior impacto (por exemplo, desenvolvendo habilidades de equipe, melhorando a performance da carga de trabalho, reduzindo custos, automatizando runbooks ou aprimorando o monitoramento). Atualize suas prioridades conforme as necessidades mudam.

Antipadrões comuns:

- Você decidiu alterar as alocações de endereços IP para suas equipes de produtos, sem consultá-las, para facilitar o gerenciamento da sua rede. Você não sabe o impacto que isso terá em suas equipes de produtos.
- Você está implementando uma nova ferramenta de desenvolvimento, mas não envolveu seus clientes internos para descobrir se ela é necessária ou se é compatível com as práticas que eles realizam.
- Você está implementando um novo sistema de monitoramento, mas não entrou em contato com seus clientes internos para descobrir se eles têm necessidades de monitoramento ou relatórios que devam ser consideradas.

Benefícios do estabelecimento desta prática recomendada: Avaliar e compreender as necessidades de clientes internos informará como você priorizará seus esforços para entregar valor empresarial.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Compreenda as necessidades empresariais: o sucesso nos negócios é possibilitado pelos objetivos e pelo entendimento compartilhados entre as partes interessadas, incluindo equipes corporativas, de desenvolvimento e de operações.
  - Analise os objetivos, as necessidades e as prioridades empresariais dos clientes internos: envolva as principais partes interessadas, incluindo as equipes corporativas, de desenvolvimento e de operações, para discutir as metas, as necessidades e as prioridades dos clientes internos. Isso garantirá que você tenha um entendimento completo do suporte às operações que é necessário para obter resultados nos negócios.
  - Estabeleça uma compreensão compartilhada: estabeleça um entendimento compartilhado das funções corporativas sobre a workload, as funções de cada uma das equipes na operação da workload e de como esses fatores apoiam seus objetivos empresariais compartilhados entre os clientes internos e externos.

Recursos

Documentos relacionados:

- AWS Well-Architected Framework Concepts – Feedback loop (Conceitos do AWS Well-Architected Framework: loop de feedback)

OPS01-BP03 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.

Antipadrões comuns:

- Você está sendo auditado e precisa fornecer prova de conformidade com a governança interna. Você não tem ideia se está em conformidade, pois nunca avaliou quais são seus requisitos de conformidade.
- Você fez um acordo que resultou em perda financeira. Você descobre que o seguro que cobriria a perda financeira dependia da sua implementação de controles de segurança específicos que não estão em vigor e são exigidos pela sua governança.
- Sua conta administrativa foi comprometida, resultando na desfiguração do site da sua empresa e na perda da confiança dos clientes. A governança interna requer o uso da autenticação multifator (MFA) para proteger as contas administrativas. Você não protegeu sua conta administrativa com MFA e está sujeito a ações disciplinares.

Benefícios do estabelecimento desta prática recomendada: Avaliar e compreender os requisitos de governança que sua organização aplica à carga de trabalho informará como você prioriza seus esforços para entregar valor empresarial.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Compreender os requisitos da governança: avalie os fatores internos da governança, como programa ou política organizacional, políticas do programa, políticas específicas de problemas ou do sistema, padrões, procedimentos, referências e diretrizes. 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.

Recursos

Documentos relacionados:

- Conformidade da Nuvem AWS

OPS01-BP04 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.

Antipadrões comuns:

- Você está sendo auditado e é solicitado a fornecer prova de conformidade com as regulamentações do setor. Você não tem ideia se está em conformidade, pois nunca avaliou quais são seus requisitos de conformidade.
- Sua conta administrativa foi comprometida, resultando no download dos dados dos clientes e na perda da confiança deles. As melhores práticas do setor exigem o uso de MFA para proteger contas administrativas. Você não protegeu sua conta administrativa com MFA e está sujeito a litígio por parte de seus clientes.

Benefícios do estabelecimento desta prática recomendada: Avaliar e compreender os requisitos de conformidade que se aplicam à sua carga de trabalho informará como você prioriza seus esforços para entregar valor empresarial.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Compreender os requisitos de conformidade: avalie os fatores externos, como requisitos de conformidade normativa e os padrões do setor, para garantir que você esteja ciente das diretrizes ou obrigações que possam exigir ou enfatizar um foco específico. Se nenhum requisito de conformidade for identificado, verifique se a auditoria devida foi aplicada à determinação.
  - Compreender os requisitos de conformidade normativa: identifique os requisitos de conformidade normativa que devem ser atendidos legalmente. Use esses requisitos para concentrar seus esforços. Os exemplos incluem as obrigações de privacidade e atos de proteção de dados.
    - Conformidade da AWS
    - Programas de conformidade da AWS
    - Notícias recentes sobre conformidade da AWS
  - Compreender os padrões e as práticas recomendadas do setor: identifique os padrões e os requisitos das práticas recomendadas do setor que se aplicam à sua workload, como o Conselho de Padrões de Segurança do Setor de Cartões de Pagamento (PCI DSS). Use esses requisitos para concentrar seus esforços.
    - Programas de conformidade da AWS
  - Compreender os requisitos internos de conformidade: identifique os requisitos de conformidade e as práticas recomendadas estabelecidas pela sua organização. Use esses requisitos para

concentrar seus esforços. Os exemplos incluem políticas de segurança da informação e padrões de classificação de dados.

Recursos

Documentos relacionados:

- Conformidade da Nuvem AWS
- Conformidade da AWS
- Notícias recentes sobre conformidade da AWS
- Programas de conformidade da AWS

OPS01-BP05 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.

O Well-Architected Framework enfatiza o aprendizado, a medição e a melhoria. Ele fornece uma abordagem consistente para avaliar arquiteturas e implementar projetos que aumentarão em escala verticalmente 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-los com as práticas recomendadas de arquitetura mais recentes da AWS, monitorar o status geral das workloads e obter insights sobre possíveis riscos.

Os clientes da AWS estão qualificados para uma revisão orientada pelo Well-Architected de suas workloads de essenciais para medir a arquitetura deles em relação às práticas recomendadas da AWS. Os clientes do Enterprise Support estão qualificados para uma Revisão de operações, projetada para ajudá-los a identificar lacunas em sua abordagem de operação na nuvem.

O envolvimento entre equipes dessas avaliações ajuda a estabelecer um entendimento comum de suas cargas de trabalho e como as funções da equipe contribuem para o sucesso. As necessidades identificadas pela avaliação podem ajudar a moldar suas prioridades.

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 suas prioridades.

Antipadrões comuns:

- Você está usando uma versão antiga de uma biblioteca de software no seu produto. Você não está ciente das atualizações de segurança na biblioteca para problemas que podem ter um impacto indesejado na carga de trabalho.
- Seu concorrente acabou de lançar uma versão do produto que lida com muitas das reclamações de seus clientes sobre seu produto. Você não priorizou a abordagem de nenhum desses problemas conhecidos.
- Os reguladores buscam empresas como a sua que não estejam em conformidade com os requisitos de conformidade normativa legais. Você não priorizou a abordagem de nenhum de seus requisitos de conformidade pendentes.

Benefícios do estabelecimento desta prática recomendada: Identificar e compreender as ameaças à sua organização e carga de trabalho permite determinar quais ameaças devem ser resolvidas, a prioridade delas e os recursos necessários para isso.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Avaliar o cenário de ameaças aos negócios: avalie as ameaças aos negócios (como concorrência, riscos e responsabilidades comerciais, riscos operacionais e ameaças à segurança das informações), para que você possa incluir o impacto dessas ameaças ao determinar onde concentrar esforços.
  - Boletins de segurança mais recentes da AWS
  - AWS Trusted Advisor
  - Manter um modelo de ameaças: estabeleça e mantenha um modelo de ameaças que identifique possíveis ameaças, mitigações planejadas e implementadas e a prioridade delas. Analise a probabilidade de as ameaças se manifestarem como incidentes, o custo de recuperação desses incidentes, o dano esperado causado e o custo para evitar esses incidentes. Revise as prioridades à medida que o conteúdo do modelo de ameaça muda.

Recursos

Documentos relacionados:

- Conformidade da Nuvem AWS
- Boletins de segurança mais recentes da AWS
- AWS Trusted Advisor

OPS01-BP06 Avalie as compensações

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 relacionais 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.

A AWS pode ajudar a educar suas equipes sobre a AWS e seus serviços para aumentar a compreensão de como suas escolhas podem ter um impacto na workload. Você deve usar os recursos fornecidos pelo AWS Support (Centro de Conhecimentos da AWS, Fóruns de discussão da AWSe AWS Support Center) e pela documentação da AWS para instruir suas equipes. Entre em contato com o AWS Support por meio do AWS Support Center para obter ajuda com relação às suas dúvidas sobre a AWS.

A AWS também compartilha as práticas recomendadas e os 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 O podcast oficial da AWS.

Antipadrões comuns:

- Você está usando um banco de dados relacional para gerenciar séries temporais e dados não relacionais. Existem opções de banco de dados otimizadas para oferecer suporte aos tipos de dados que você está usando, mas você não tem conhecimento dos benefícios, pois não avaliou as compensações entre soluções.
- Seus investidores solicitam que você demonstre conformidade com os Padrões de segurança de dados do setor de cartões de pagamento (PCI DSS). Você não considera as compensações entre atender à solicitação deles e continuar com seus esforços de desenvolvimento atuais. Em vez disso, prossiga com seus esforços de desenvolvimento sem demonstrar conformidade. Seus

investidores interrompem o suporte da sua empresa devido a preocupações com a segurança da sua plataforma e com os investimentos deles.

Benefícios do estabelecimento desta prática recomendada: Entender as implicações e as consequências de suas escolhas permite que você priorize suas opções.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Avaliar as compensações: avalie o impacto das compensações entre as partes interessadas concorrentes para ajudar a tomar decisões embasadas ao determinar onde concentrar esforços. Por exemplo, a aceleração da velocidade de introdução no mercado de novos recursos pode ser enfatizada sobre a otimização de custos.
- A AWS pode ajudar a educar suas equipes sobre a AWS e seus serviços para aumentar a compreensão de como suas escolhas podem ter um impacto na workload. Use os recursos fornecidos pelo AWS Support (Centro de Conhecimentos da AWS, Fóruns de discussão da AWS e AWS Support Center) e pela documentação da AWS para instruir suas equipes. Entre em contato com o AWS Support por meio do AWS Support Center para obter ajuda com relação às suas dúvidas sobre a AWS.
- A AWS também compartilha as práticas recomendadas e os padrões que aprendemos durante a operação da AWS na Amazon Builders' Library. Uma grande variedade de outras informações úteis está disponível no Blog da AWS e no podcast oficial da AWS.

Recursos

Documentos relacionados:

- Blog da AWS
- Conformidade da Nuvem AWS
- Fóruns de discussão da AWS
- documentação da AWS
- Centro de Conhecimentos da AWS
- AWS Support
- AWS Support Center

- Amazon Builders' Library
- O podcast oficial da AWS

OPS01-BP07 Gerenciar os benefícios e os 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.

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. Atualize suas prioridades conforme as necessidades mudam

Antipadrões comuns:

- Um de seus desenvolvedores encontrou na Internet, uma biblioteca que faz tudo o que você precisa, e você decidiu incluí-la. Você não avaliou os riscos de adoção dessa biblioteca de uma origem desconhecida e não sabe se ela contém vulnerabilidades ou código mal-intencionado.
- Você decidiu desenvolver e implantar um novo recurso em vez de corrigir um problema existente. Você não avaliou os riscos de continuar com o problema até que o recurso seja implantado e não sabe qual será o impacto nos seus clientes.
- Você decidiu não implantar um recurso solicitado frequentemente pelos clientes devido a preocupações não especificadas da sua equipe de conformidade.

Benefícios do estabelecimento desta prática recomendada: Identificar os benefícios disponíveis das suas escolhas e estar ciente dos riscos para a sua organização permite que você tome decisões bem embasadas.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Gerenciar os benefícios e os riscos: equilibre os benefícios das decisões em relação aos riscos envolvidos.

- Identificar os benefícios: identifique os benefícios com base nas metas, necessidades e prioridades da empresa. Os exemplos incluem tempo de colocação no mercado, segurança, confiabilidade, performance e custo.
- Identificar os riscos: identifique os riscos com base nas metas, necessidades e prioridades da empresa. Os exemplos incluem tempo de colocação no mercado, segurança, confiabilidade, performance e custo.
- Avaliar os benefícios em relação aos riscos e tomar decisões embasadas: determine o impacto dos benefícios e dos riscos com base nas metas, necessidades e prioridades das principais partes interessadas, incluindo os negócios, o desenvolvimento e as operações. Avalie o valor do benefício em relação à probabilidade de realização do risco e o custo do seu impacto. Por exemplo, enfatizar a velocidade de entrada no mercado em vez da confiabilidade pode oferecer vantagem competitiva. No entanto, isso pode resultar em tempo de atividade reduzido se houver problemas de confiabilidade.

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.

Práticas recomendadas

- OPS02-BP01 Recursos com proprietários identificados
- OPS02-BP02 Processos e procedimentos com proprietários identificados
- OPS02-BP03 Atividades de operações com proprietários identificados responsáveis pela performance
- OPS02-BP04 Os membros da equipe sabem pelo que são responsáveis
- OPS02-BP05 Existem mecanismos para identificar a responsabilidade e a propriedade
- OPS02-BP06 Mecanismos existem para solicitar adições, alterações e exceções
- OPS02-BP07 As responsabilidades entre as equipes são predefinidas ou negociadas

OPS02-BP01 Recursos com proprietários identificados

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.

Benefícios do estabelecimento desta prática recomendada: Entender a propriedade identifica quem pode aprovar melhorias, implementar essas melhorias ou ambos.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Recursos com proprietários identificados: defina o que significa propriedade para os casos de uso de recursos em seu ambiente. Especifique e registre proprietários para recursos, incluindo pelo menos nome, informações de contato, organização e equipe. Armazene informações de propriedade de recursos com recursos usando metadados, como tags ou grupos de recursos. Use o AWS Organizations para estruturar contas e implementar políticas a fim de garantir que as informações de propriedade e de contatos sejam capturadas.
  - Definir formas de propriedade e como elas são atribuídas: propriedade pode ter várias definições em sua organização, com diferentes casos de uso. É possível definir um proprietário da workload como o indivíduo que possui o risco e a responsabilidade pela operação de uma workload e que, em última análise, tem autoridade para tomar decisões sobre a workload. Você pode querer definir a propriedade em termos de responsabilidade financeira ou administrativa, em que a propriedade passa para uma organização pai. Um desenvolvedor pode ser o proprietário do ambiente de desenvolvimento e ser responsável pelos incidentes que a operação causa. O líder do produto pode ser responsável pelos custos financeiros associados à operação dos ambientes de desenvolvimento.
  - Definir quem possui uma organização, conta, coleção de recursos ou componentes individuais: defina e registre a propriedade em um local acessível e organizado para apoiar a descoberta de forma apropriada. Atualize definições e detalhes de propriedade à medida que eles mudarem.
  - Capturar a propriedade dos recursos nos metadados: capture a propriedade dos recursos usando metadados, como tags ou grupos de recursos, especificando as informações de contato e de propriedade. Use o AWS Organizations para estruturar as contas e garantir que as informações de contato e de propriedade sejam capturadas.

OPS02-BP02 Processos e procedimentos com proprietários identificados

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.

Benefícios do estabelecimento desta prática recomendada: Entender a propriedade identifica quem pode aprovar melhorias, implementar essas melhorias ou ambos.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Processos e procedimentos com proprietários identificados responsáveis pela sua definição: capture os processos e procedimentos usados em seu ambiente e o indivíduo ou a equipe responsável pela sua definição.
  - Identifique processos e procedimentos: identifique as atividades de operações realizadas para dar suporte às suas workloads. Documente essas atividades em um local que possa ser localizado.
  - Defina quem é o proprietário de um processo ou procedimento: identifique exclusivamente o indivíduo ou a equipe responsável pela especificação de uma atividade. Eles são responsáveis por garantir que ela possa ser executada com êxito por um membro da equipe devidamente qualificado com as permissões, as ferramentas e o acesso corretos. Se houver problemas com a execução dessa atividade, os membros da equipe que a executam serão responsáveis por fornecer os comentários detalhados necessários para que a atividade seja melhorada.
  - Capture a propriedade de artefato de atividades nos metadados: os procedimentos automatizados em serviços como o AWS Systems Manager, por meio de documentos, e o AWS Lambda, como funções, são compatíveis com a captura de informações de metadados como tags. Capture a propriedade de recursos usando tags ou grupos de recursos, especificando propriedade e informações de contato. Use o AWS Organizations para criar políticas de marcação e garantir que as informações de propriedade e de contato sejam capturadas.

OPS02-BP03 Atividades de operações com proprietários identificados 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.

Benefícios do estabelecimento desta prática recomendada: Entender quem é responsável por realizar uma atividade informa a quem notificar quando uma ação é necessária e quem executará a ação, validará o resultado e fornecerá feedback ao proprietário da atividade.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Atividades de operações com proprietários identificados responsáveis por sua performance: capture a responsabilidade por executar processos e procedimentos usados em seu ambiente.
  - Identificar processos e procedimentos: identifique as atividades de operações realizadas para dar suporte às suas workloads. Documente essas atividades em um local que possa ser localizado.
  - Definir quem é responsável por executar cada atividade: identifique a equipe responsável por uma atividade. Certifique-se de que eles tenham os detalhes da atividade e as habilidades necessárias e as permissões, as ferramentas e o acesso corretos para realizar a atividade. Eles devem compreender a condição sob a qual ela deve ser executada (por exemplo, em um evento ou programação). Torne essas informações detectáveis para que os membros da sua organização possam identificar com quem precisam entrar em contato, equipe ou indivíduo, para necessidades específicas.

OPS02-BP04 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.

Benefícios do estabelecimento desta prática recomendada: Entender suas responsabilidades informa as decisões que você toma, as ações que você realiza e suas atividades de entrega aos proprietários apropriados.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Garantir que os membros da equipe compreendam suas funções e responsabilidades: identifique as funções e responsabilidades dos membros da equipe e garanta que eles compreendam as

expectativas da função que exercem. Torne essas informações detectáveis para que os membros da sua organização possam identificar com quem precisam entrar em contato, equipe ou indivíduo, para necessidades específicas.

OPS02-BP05 Existem mecanismos para identificar a responsabilidade e a 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.

Benefícios do estabelecimento desta prática recomendada: Entender quem tem responsabilidade ou propriedade permite que você entre em contato com a equipe ou o membro da equipe apropriado para fazer uma solicitação ou a transição de uma tarefa. Ter uma pessoa identificada que tenha a autoridade para atribuir responsabilidade ou propriedade ou planejar atender às necessidades, reduz o risco de inação, além de não ser preciso lidar com isso.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Mecanismos existentes para identificar a responsabilidade e a propriedade: forneça mecanismos acessíveis para que os membros da sua organização descubram e identifiquem a propriedade e a responsabilidade. Esses mecanismos permitirão que eles identifiquem com quem entrar em contato, equipe ou indivíduo, em caso de necessidades específicas.

OPS02-BP06 Mecanismos existem 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.

Benefícios do estabelecimento desta prática recomendada: É essencial que existam mecanismos para solicitar adições, alterações e exceções para apoiar as atividades das equipes. Sem essa opção, o estado atual se torna uma restrição de inovação.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Mecanismos existem para solicitar adições, alterações e exceções: quando os padrões são rígidos, a inovação é limitada. Forneça mecanismos para que os membros da sua organização

façam solicitações aos proprietários de processos, procedimentos e recursos para atender às necessidades comerciais deles.

OPS02-BP07 As responsabilidades entre as equipes são predefinidas ou negociadas

Tenha acordos definidos ou negociados entre as equipes que descrevam como elas trabalham e oferecem suporte umas às outras (por exemplo, tempos de resposta, objetivos de nível de serviço ou Acordos 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.

Quando a responsabilidade e a propriedade não foram definidas ou são desconhecidas, você corre o risco de não abordar as atividades necessárias em tempo hábil e de despender esforços redundantes e possivelmente conflitantes para atender a essas necessidades.

Benefícios do estabelecimento desta prática recomendada: Estabelecer as responsabilidades entre as equipes, os objetivos e os métodos de comunicação das necessidades facilita o fluxo de solicitações e ajuda a garantir que as informações necessárias sejam fornecidas. Isso reduz o atraso introduzido pelas tarefas de transição entre equipes e ajuda a apoiar a obtenção de resultados empresariais.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Responsabilidades entre as equipes são predefinidas e negociadas: especificar os métodos pelos quais as equipes interagem e as informações necessárias para que ofereçam suporte umas para as outras ajuda a minimizar o atraso introduzido à medida que as solicitações são analisadas e esclarecidas iterativamente. Ter contratos específicos que definem expectativas (por exemplo, tempo de resposta ou de atendimento) permite que as equipes criem planos e recursos eficazes de modo adequado.

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.

Práticas recomendadas

- OPS03-BP01 Patrocínio executivo

- OPS03-BP02 Os membros da equipe estão capacitados para executar ações quando os resultados estão em risco.
- OPS03-BP03 Incentivo ao escalonamento
- OPS03-BP04 Comunicações oportunas, claras e acionáveis
- OPS03-BP05 Incentivo à experimentação
- OPS03-BP06 Os membros da equipe estão capacitados e são incentivados a manter e a aumentar seus conjuntos de habilidades.
- OPS03-BP07 Fornecer recursos adequados às equipes
- OPS03-BP08 Opiniões diversas são incentivadas e procuradas dentro e entre equipes

OPS03-BP01 Patrocínio 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

Benefícios do estabelecimento desta prática recomendada: A liderança engajada, as expectativas comunicadas claramente e as metas compartilhadas garantem que os membros da equipe saibam o que se espera deles. A avaliação do sucesso possibilita a identificação de barreiras para o sucesso, para que elas possam ser abordadas por meio da intervenção do patrocinador ou dos representantes dele.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Patrocínio 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
  - Definir as expectativas: defina e publique metas para suas organizações, incluindo como elas serão medidas.
  - Monitorar a concretização das metas: meça regularmente a concretização incremental das metas e compartilhe os resultados para que medidas adequadas possam ser tomadas se os resultados estiverem em risco.
  - Fornecer os recursos necessários para realizar suas metas: analise regularmente se os recursos ainda são apropriados ou se recursos adicionais são necessários com base em novas informações, alterações nas metas, responsabilidades ou ambiente da empresa.

- Defender suas equipes: mantenha o envolvimento com suas equipes para compreender a performance delas e se estão sendo afetadas por fatores externos. Quando suas equipes forem afetadas por fatores externos, reavalie metas e ajuste os objetivos conforme apropriado. Identifique os obstáculos que estão impedindo o progresso das suas equipes. Aja em nome das suas equipes para ajudar a resolver obstáculos e eliminar obrigações desnecessárias.
- Motivar a adoção de práticas recomendadas: confirme as práticas recomendadas que oferecem benefícios quantificáveis e reconheça quem as cria e as adota. Incentive ainda mais a adoção para ampliar os benefícios obtidos.
- Motivar a evolução de suas equipes: crie uma cultura de melhoria contínua. Incentive o crescimento e o desenvolvimento pessoal e organizacional. Forneça metas de longo prazo pelas quais se esforçar que exigirão conquistas incrementais ao longo do tempo. Ajuste essa visão para complementar necessidades, metas de negócios e ambiente de negócios à medida que eles mudarem.

OPS03-BP02 Os membros da equipe estão capacitados para 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.

Benefícios do estabelecimento desta prática recomendada: Ao testar e validar alterações antecipadamente, você pode resolver problemas com custos reduzidos e limitar o impacto sobre seus clientes. Ao testar antes da implantação, você reduz a possibilidade de erros.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Os membros da equipe estão capacitados para executar ações quando os resultados estão em risco: forneça aos membros da equipe as permissões, as ferramentas e a oportunidade de praticar as habilidades necessárias para responderem com eficácia.
  - Fornecer aos membros da equipe a oportunidade de praticar as habilidades necessárias para responder: forneça ambientes alternativos e seguros em que os processos e os procedimentos possam ser testados e treinados com segurança. Realizar dias de jogos para permitir que os membros da equipe adquiram experiência para responder a incidentes reais em ambientes simulados e seguros.

- Definir e confirmar a autoridade dos membros da equipe para executar ações: defina especificamente a autoridade dos membros da equipe para executar ações por meio da atribuição de permissões e acesso às workloads e aos componentes aos quais oferecem suporte. Reconheça que eles estão capacitados a executar ações quando os resultados estão em risco.

OPS03-BP03 Incentivo ao 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.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Incentivar o escalonamento antecipado e frequente: reconheça de forma organizacional que o escalonamento antecipado e frequente é a prática recomendada. Reconheça e aceite de maneira organizacional que os escalonamentos podem ser infundados e que é melhor ter a oportunidade de evitar um incidente do que perder essa oportunidade ao não escalonar.
  - Ter um mecanismo para o escalonamento: tenha procedimentos documentados que definem quando e como o escalonamento deve ocorrer. Documente a série de pessoas com autoridade crescente para tomar medidas ou aprovar ações e as informações de contato delas. O escalonamento deve continuar até que o membro da equipe esteja satisfeito por ter transmitido o risco a alguém capaz de lidar com ele ou tenha entrado em contato com a pessoa que detém o risco e a responsabilidade pela operação da workload. É essa pessoa que, em última análise, tem todas as decisões com relação à carga de trabalho. Os escalonamentos devem incluir a natureza do risco, a criticidade da carga de trabalho, quem é afetado, qual é o impacto e a urgência, ou seja, quando é o impacto esperado.
  - Proteger os funcionários que usam o escalonamento: tenha uma política que proteja os membros da equipe contra retaliações se fizerem um escalonamento em relação a um tomador de decisão ou parte interessada não responsivo. Tenha mecanismos implementados para identificar se isso está ocorrendo e responder de maneira adequada.

OPS03-BP04 Comunicações 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.

Eventos planejados podem ser registrados em um calendário de alterações ou programação de manutenção para que os membros da equipe possam identificar quais atividades estão pendentes.

Na AWS, Calendário de alterações do AWS Systems Manager pode ser usado para registrar esses detalhes. Ele oferece suporte a verificações programáticas do status do calendário para determinar se o calendário está aberto ou fechado para atividades em um determinado momento. As atividades de operações podem ser planejadas em torno de períodos específicos e aprovados reservados para atividades que potencialmente causam interrupções. As janelas de manutenção do AWS Systems Manager permitem programar atividades com base em instâncias e outros recursos com suporte para automatizar as atividades e torná-las detectáveis.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- As comunicações são oportunas, claras e acionáveis: foram implementados mecanismos para fornecer notificações de riscos ou de eventos planejados de maneira clara e acionável com aviso prévio em tempo suficiente para permitir respostas apropriadas.
  - Documentar atividades planejadas em um calendário de alterações e fornecer notificações: fornecer uma fonte acessível de informações em que os eventos planejados possam ser descobertos. Forneça notificações de eventos planejados oriundos do mesmo sistema.
  - Rastrear eventos e atividades que podem ter impacto sobre a workload: monitore as notificações de vulnerabilidade e informações sobre patches para compreender as vulnerabilidades de riscos reais e potenciais associados aos componentes da workload. Forneça uma notificação aos membros da equipe para que eles possam executar ações.

Recursos

Documentos relacionados:

- Calendário de alterações do AWS Systems Manager
- AWS Systems Manager Maintenance Windows

OPS03-BP05 Incentivo à 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.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Incentivo à experimentação: incentive a experimentação para apoiar o aprendizado e a inovação.
  - Experimentar com uma variedade de tecnologias: incentive a experimentação com tecnologias que possam ser aplicáveis agora ou no futuro para a obtenção dos resultados empresariais. Esse conhecimento pode informar inovações futuras.
  - Experimentar com uma meta em mente: incentive a experimentação com metas específicas para os membros da equipe atingirem ou com tecnologias que sejam aplicáveis em um futuro próximo. Esse conhecimento pode informar sua inovação.
  - Fornecer tempo estruturado para experimentos: dedique momentos específicos em que os membros da equipe possam ficar livres das responsabilidades normais e concentrarem-se em experimentos.
  - Fornecer os recursos para apoiar a experimentação: forneça os recursos necessários para a realização de experimentos (por exemplo, software ou recursos de nuvem).
  - Reconhecer o sucesso: reconheça o valor gerado pela experimentação. Compreenda que os experimentos com resultados indesejados são bem-sucedidos e identificaram um caminho que não levará ao sucesso. Os membros da equipe não são punidos por resultados indesejados de experimentos.

OPS03-BP06 Os membros da equipe estão capacitados e são incentivados a manter e a aumentar seus 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 quando você perde membros da equipe qualificados e experientes com conhecimento institucional. Reserve tempo estruturado e dedicado para o aprendizado.

A AWS fornece recursos, incluindo o Centro de recursos de conceitos básicos da AWS, Blogs da AWS, AWS Online Tech Talks, Eventos e webinars da AWSe os Laboratórios do AWS Well-Architected, que fornecem orientações, exemplos e demonstrações detalhadas para educar suas equipes.

A AWS também compartilha as práticas recomendadas e os padrões que aprendemos durante a operação da AWS na Amazon Builders' Library e uma grande variedade de outros materiais educacionais úteis por meio do Blog da AWS e O podcast oficial da AWS.

Você deve aproveitar os recursos educacionais fornecidos pela AWS, como os laboratórios do AWS Well-Architected, AWS Support (Centro de Conhecimentos da AWS, Fóruns de discussão da AWSe aos AWS Support Center) e Documentação da AWS para instruir suas equipes. Entre em contato com o AWS Support por meio do AWS Support Center para obter ajuda com relação às suas dúvidas sobre a AWS.

Treinamento da AWS and Certification fornece alguns treinamentos gratuitos por meio de cursos digitais autoguiados sobre os conceitos básicos da AWS. Também é possível inscrever-se em treinamento administrado por instrutor para oferecer suporte adicional ao desenvolvimento das habilidades em AWS de suas equipes.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Os membros da equipe estão capacitados e são incentivados a manter e a ampliar seus conjuntos de habilidades: para adotar novas tecnologias e oferecer suporte às mudanças na demanda e nas responsabilidades de suporte às workloads, é necessário treinamento contínuo.
  - Fornecer recursos para treinamento: forneça tempo estruturado dedicado, acesso ao material de treinamento, recursos de laboratório e suporte à participação em conferências e organizações profissionais que fornecem oportunidades para aprendizado para instrutores e colegas. Forneça aos membros da equipe júnior acesso aos membros da equipe sênior como mentores ou permita que eles sigam o trabalho deles e sejam expostos aos métodos e às habilidades que

têm. Incentive o aprendizado sobre conteúdo não diretamente relacionado ao trabalho para ter uma perspectiva mais ampla.

- Treinamento da equipe e envolvimento entre equipes: planeje as necessidades de treinamento contínuo dos membros da equipe. Ofereça oportunidades para que os membros da equipe se juntem a outras equipes (temporária ou permanentemente) para compartilhar habilidades e melhores práticas que beneficiam toda a organização
- Oferecer suporte à busca e à manutenção de certificações do setor: ofereça suporte à aquisição e manutenção de certificações do setor que validam o aprendizado e reconheça as conquistas dos membros da equipe.

Recursos

Documentos relacionados:

- Centro de recursos de conceitos básicos da AWS
- Blogs da AWS
- Conformidade da Nuvem AWS
- Fóruns de discussão da AWS
- Documentação da AWS
- AWS Online Tech Talks
- Eventos e webinars da AWS
- Centro de Conhecimentos da AWS
- AWS Support
- Treinamento da AWS and Certification
- Laboratórios do AWS Well-Architected,
- Amazon Builders' Library
- O podcast oficial da AWS.

OPS03-BP07 Fornecer recursos adequados às equipes

Mantenha a capacidade dos membros da equipe e forneça ferramentas e recursos para dar suporte às necessidades da workload. A sobrecarga de membros da equipe aumenta o risco de incidentes resultantes de erros humanos. Os investimentos em ferramentas e em recursos (por exemplo, o

fornecimento de automação para atividades executadas com frequência) podem escalar a eficácia da equipe, permitindo que ela ofereça suporte a atividades adicionais.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Fornecer recursos adequados às equipes: compreenda o sucesso das equipes e os fatores que contribuem para o sucesso ou para o insucesso. Aja para apoiar equipes com os recursos apropriados.
  - Compreender a performance da equipe: meça a aquisição de resultados operacionais e o desenvolvimento de ativos realizados pela equipe. Acompanhe as alterações na saída e na taxa de erros ao longo do tempo. Envolva-se com as equipes para compreender os desafios relacionados ao trabalho que as afetam (por exemplo, aumento de responsabilidades, mudanças na tecnologia, perda de pessoal ou aumento de clientes atendidos pelo suporte).
  - Compreender os impactos na performance das equipes: mantenha-se engajado com as equipes para entender como elas estão desempenhando e se há fatores externos que as afetam. Quando suas equipes forem afetadas por fatores externos, reavalie metas e ajuste os objetivos conforme apropriado. Identifique os obstáculos que estão impedindo o progresso das suas equipes. Aja em nome das suas equipes para ajudar a resolver obstáculos e eliminar obrigações desnecessárias.
  - Fornecer os recursos necessários para as equipes serem bem-sucedidas: analise regularmente se os recursos ainda são adequados, ou se são necessários recursos adicionais, e faça os ajustes apropriados para oferecer suporte às equipes.

OPS03-BP08 Opiniões diversas 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.

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.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Buscar opiniões e perspectivas diversas: incentive contribuições de todos. Ouça os grupos de pequena representação. Alterne as funções e as responsabilidades em reuniões.
  - Expandir funções e responsabilidades: ofereça oportunidade para que os membros da equipe assumam funções que não poderiam assumir de outra forma. Eles ganharão experiência e perspectiva com a função e com as interações com novos membros da equipe com os quais não interagiriam de outra forma. Eles levarão a experiência e perspectiva deles para a nova função e para os membros da equipe com os quais interagirem. Conforme aumenta a perspectiva, mais oportunidades de negócios podem surgir ou novas oportunidades de melhoria podem ser identificadas. Faça com que os membros de uma equipe se revezem em tarefas comuns que outras pessoas normalmente executam para compreender as demandas e o impacto de realizá-las.
  - Fornecer um ambiente seguro e acolhedor: implante políticas e controles que protejam a segurança física e mental dos membros da equipe na organização. Os membros da equipe devem poder interagir sem medo de represálias. Quando os membros da equipe se sentem seguros e bem-vindos, eles provavelmente estão envolvidos e são produtivos. Quanto mais diversificada sua organização, melhor será o entendimento das pessoas que você apoia, incluindo seus clientes. Quando os membros da equipe estiverem confortáveis, sentirem-se à vontade para falar e confiarem que serão ouvidos, será mais provável que eles dividam ideias valiosas (por exemplo, oportunidades de marketing, necessidades de acessibilidade, segmentos de mercado não atendidos, riscos não reconhecidos no seu ambiente).
  - Permitir que os membros da equipe participem plenamente: forneça os recursos necessários para que os funcionários participem totalmente de todas as atividades relacionadas ao trabalho. Os membros da equipe que enfrentam desafios diários desenvolveram habilidades para contornar esses desafios. Essas habilidades desenvolvidas exclusivamente podem oferecer benefícios significativos para a sua organização. O apoio aos membros da equipe com as acomodações necessárias aumentará os benefícios que você poderá receber das contribuições deles.

Preparar

Perguntas

- OPS 4 Como você projeta sua carga de trabalho para entender o estado dela?
- OPS 5 Como você reduz defeitos, facilita a correção e melhora o fluxo na produção?

- OPS 6 Como você reduz os riscos de implantação?
- OPS 7 Como você sabe que está pronto para oferecer suporte a uma carga de trabalho?

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.

Práticas recomendadas

- OPS04-BP01 Implementar a telemetria de aplicações
- OPS04-BP02 Implementar e configurar a telemetria da workload
- OPS04-BP03 Implementar a telemetria de atividades dos usuários
- OPS04-BP04 Implementar a telemetria de dependências
- OPS04-BP05 Implementar a capacidade de rastreamento das transações

OPS04-BP01 Implementar a telemetria de aplicações

A telemetria de aplicações é a base para a observabilidade da workload. A aplicação deve emitir telemetria que forneça insight do estado da aplicação e da aquisição dos resultados da empresa. Da solução de problemas à medição do impacto de um novo recurso, a telemetria de aplicações informa a maneira como você cria, opera e evolui a workload.

A telemetria de aplicações consiste em métricas e logs. As métricas são informações de diagnóstico, como seu pulso ou temperatura. As métricas são usadas coletivamente para descrever o estado de uma aplicação. A coleta das métricas ao longo do tempo pode ser usada para desenvolver linhas de base e detectar anomalias. Os logs são mensagens que a aplicação envia sobre seu estado interno ou os eventos que ocorrem. Códigos de erros, identificadores de transações e ações dos usuários são exemplos dos eventos registrados em log.

Resultado desejado:

- A aplicação emite métricas e logs que fornecem insights da integridade e da aquisição de resultados dos negócios.
- As métricas e logs são armazenados centralmente para todas as aplicações na workload.

Antipadrões comuns:

- Seu aplicativo não emite telemetria. Você é forçado a contar com seus clientes para informar quando algo está errado.
- Um cliente relatou que seu aplicativo não responde. Você não tem telemetria e não consegue confirmar se o problema existe ou caracterizar o problema sem usar o aplicativo para entender a experiência atual do usuário.

Benefícios do estabelecimento desta prática recomendada:

- É possível compreender a integridade das aplicações, a experiência dos usuários e a aquisição dos resultados.
- Reagir rapidamente às mudanças da integridade das aplicações.
- Desenvolver a partir das tendências da integridade das aplicações.
- Tomar decisões embasadas sobre como melhorar as aplicações.
- E detectar e resolver problemas das aplicações mais rapidamente.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

A implementação da telemetria de aplicações consiste em três etapas: a identificação de um local para armazenar a telemetria, a identificação da telemetria que descreve o estado das aplicações e a instrumentação das aplicações para emitirem telemetria.

Como exemplo, uma empresa de comércio eletrônico tem uma arquitetura baseada em microsserviços. Como parte do processo de design dessa arquitetura, a empresa identificou a telemetria de aplicações que a ajudaria a entender o estado de cada microsserviço. Por exemplo, o serviço de carrinho do usuário emite telemetria sobre eventos, como adição ao carrinho, abandono do carrinho e o tempo levado para adicionar um item ao carrinho. Todos os microsserviços registram erros, avisos e informações sobre as transações em log. A telemetria é enviada para o Amazon CloudWatch para armazenamento e análise.

Etapas da implementação

A primeira etapa é identificar um local central para armazenamento da telemetria para as aplicações da workload. Se você ainda não tiver uma plataforma, o Amazon CloudWatch fornecerá a coleta da telemetria, os painéis, a análise e os recursos para a geração de eventos.

Para identificar a telemetria necessária, comece com as seguintes perguntas:

- Minha aplicação é íntegra?
- Minha aplicação está trazendo resultados para os negócios.

A aplicação deve emitir logs e métricas que respondam coletivamente a essas perguntas. Se não for possível responder a essas perguntas com a telemetria de aplicações existentes, trabalhe com as partes interessadas da empresa e da engenharia para criarem uma lista de telemetria que possa. É possível solicitar consultoria técnica especializada da equipe da Conta da AWS ao identificar e desenvolver nova telemetria de aplicações.

Quando a telemetria adicional de aplicações estiver identificada, trabalhe com as partes interessadas da engenharia para instrumentar as aplicações. O AWS Distro for Open Telemetry fornece bibliotecas de APIs e agentes que coletam telemetria de aplicações. Este exemplo demonstra como instrumentar uma aplicação JavaScript com métricas personalizadas..

Os clientes que quiserem compreender os serviços de observabilidade oferecidos pela AWS podem trabalhar com o Um workshop de observabilidade por conta própria ou solicitar suporte da equipe da Conta da AWS para receberem orientações. Esse workshop fornece orientações sobre as soluções de observabilidade da AWS e exemplos práticos de como elas são usadas.

Para mergulhar mais profundamente na telemetria de aplicações leia o artigo instrumentação de sistemas distribuídos para visibilidade operacional na Amazon Builder's Library. Ele explica como a Amazon instrumenta as aplicações e pode servir como um guia para o desenvolvimento de suas próprias diretrizes de instrumentação.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

the section called “OPS04-BP02 Implementar e configurar a telemetria da workload” A telemetria de aplicações é um componente da telemetria de workload. Para compreender a integridade da workload geral, entenda a integridade das aplicações individuais que compõem a workload.

the section called “OPS04-BP03 Implementar a telemetria de atividades dos usuários” A telemetria das atividades dos usuários geralmente é um subconjunto da telemetria de aplicações. As atividades

dos usuários, como eventos de adições ao carrinho, cliques em streams ou transações concluídas fornecem insight da experiência do usuário.

the section called “OPS04-BP04 Implementar a telemetria de dependências” As verificações de dependências estão relacionadas à telemetria de aplicações e podem ser instrumentadas nas aplicações. Se a aplicação contar com dependências externas, como o DNS ou um banco de dados, a aplicação poderá emitir métricas e logs sobre a acessibilidade, os tempos limite e outros eventos.

the section called “OPS04-BP05 Implementar a capacidade de rastreamento das transações” O rastreamento das transações em uma workload requer que cada aplicação emita informações sobre como ela processa eventos compartilhados. A forma como as aplicações individuais tratam esses eventos é emitida por meio da telemetria de aplicações.

the section called “OPS08-BP02 Definir as métricas da workload” As métricas da workload são os principais indicadores da integridade da workload. As métricas principais das aplicações são parte das métricas da workload.

Documentos relacionados:

- AWS Builders' Library: Como instrumentar sistemas distribuídos para obter observabilidade operacional
- AWS Distro for OpenTelemetry
- AWS Whitepaper Well-Architected Operational Excellence: Design Telemetry (Excelência operacional do Well-Architected: Design de telemetria)
- Creating metrics from log events using filters (Criação de métricas de eventos de logs usando filtros)
- Implementing Logging and Monitoring with Amazon CloudWatch (Implementação de registro em log e monitoramento com o Amazon CloudWatch)
- Monitoring application health and performance with AWS Distro for OpenTelemetry (Monitoramento da integridade e da performance das aplicações com o AWS Distro for OpenTelemetry)
- New: How to better monitor your custom application metrics using Amazon CloudWatch Agent (Novidade: Como monitorar melhor as métricas de aplicações personalizadas usando o agente do CloudWatch)
- Observability at AWS (Observabilidade na AWS)
- Scenario – Publish metrics to CloudWatch (Cenário: Publicar métricas no CloudWatch)
- Comece a criar: Como monitorar suas aplicações com eficácia
- Using CloudWatch with an AWS SDK (Usar o CloudWatch com um AWS SDK)

Vídeos relacionados:

- AWS re:Invent 2021: Observability the open-source way (re:Invent da AWS de 2021: observabilidade por código aberto)
- Collect Metrics and Logs from Amazon EC2 instances with the CloudWatch Agent (Coletar métricas e logs das instâncias do Amazon EC2 com o agente do CloudWatch)
- How to Easily Setup Application Monitoring for Your AWS Workloads - AWS Online Tech Talks (Como configurar facilmente o monitoramento de aplicações para as workloads da AWS: AWS Online Tech Talks)
- Mastering Observability of Your Serverless Applications - AWS Online Tech Talks (Domínio da observabilidade de aplicações de tecnologia sem servidor: AWS Online Tech Talks)
- Open Source Observability with AWS - AWS Virtual Workshop (Observabilidade de código aberto com a AWS: Workshop virtual da AWS)

Exemplos relacionados:

- Recursos de exemplo de registro em log e monitoramento da AWS
- AWS Solution: Amazon CloudWatch Monitoring Framework (Solução da AWS: Framework de monitoramento do AWS CloudWatch)
- AWS Solution: Centralized Logging (Solução da AWS: Registro em log centralizado)
- Um workshop de observabilidade

OPS04-BP02 Implementar e configurar a telemetria da workload

Projete e configure a workload para emitir informações sobre o estado interno e o status atual, como o volume de chamadas da API, os códigos de status HTTP e os eventos de escalabilidade. Use essas informações para auxiliá-lo na determinação de quando uma resposta é necessária.

Use um serviço, como o Amazon CloudWatch para agregar logs e métricas de componentes de carga de trabalho (por exemplo, logs de API do AWS CloudTrail, métricas do AWS Lambda, logs de fluxo da Amazon VPC e aos outros serviços).

Antipadrões comuns:

- Seus clientes estão reclamando sobre performance insatisfatória. Não há alterações recentes em seu aplicativo e, portanto, você suspeita de um problema com um componente de carga

de trabalho. Você não tem telemetria para analisar e determinar quais componentes estão contribuindo para a performance insatisfatória.

- Seu aplicativo está inacessível. Você não tem a telemetria para determinar se é um problema de rede.

Benefícios do estabelecimento desta prática recomendada: Entender o que está acontecendo dentro da sua carga de trabalho permite que você responda, se necessário.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Implementar telemetria de log e de métricas: prepare a workload para emitir informações sobre o estado interno, o status e a obtenção de resultados dos negócios. Use essas informações para determinar quando uma resposta é necessária.
  - Gaining better observability of your VMs with Amazon CloudWatch - AWS Online Tech Talks (Como obter melhor observabilidade das VMs com o Amazon CloudWatch: AWS Online Tech Talks)
  - How Amazon CloudWatch works (Como funciona o Amazon CloudWatch)
  - What is Amazon CloudWatch (O que é o Amazon CloudWatch?)
  - Using Amazon CloudWatch metrics (Uso de métrica do Amazon CloudWatch)
  - What is Amazon CloudWatch Logs? (O que é o Amazon CloudWatch?)
    - Implementar e configurar telemetria na workload: projete e configure a workload para emitir informações sobre o estado interno e o status atual (como volume de chamadas da API, códigos de status HTTP e eventos de escalabilidade).
      - Amazon CloudWatch metrics and dimensions reference (Referência de métricas e de dimensões do Amazon CloudWatch)
      - AWS CloudTrail
      - What Is AWS CloudTrail? (O que é o Amazon CloudTrail?)
      - Logs de fluxo da VPC

Recursos

Documentos relacionados:

- AWS CloudTrail

- Documentação do Amazon CloudWatch
- Amazon CloudWatch metrics and dimensions reference (Referência de métricas e de dimensões do Amazon CloudWatch)
- How Amazon CloudWatch works (Como funciona o Amazon CloudWatch)
- Using Amazon CloudWatch metrics (Uso de métrica do Amazon CloudWatch)
- Logs de fluxo da VPC
- What Is AWS CloudTrail? (O que é o Amazon CloudTrail?)
- What is Amazon CloudWatch Logs? (O que é o Amazon CloudWatch?)
- What is Amazon CloudWatch (O que é o Amazon CloudWatch?)

Vídeos relacionados:

- Application Performance Management on AWS (Gerenciamento da performance de aplicações na AWS)
- Gaining Better Observability of Your VMs with Amazon CloudWatch (Como obter melhor observabilidade de suas VMs com o Amazon CloudWatch)
- Gaining better observability of your VMs with Amazon CloudWatch - AWS Online Tech Talks (Como obter melhor observabilidade das VMs com o Amazon CloudWatch: AWS Online Tech Talks)

OPS04-BP03 Implementar a telemetria de atividades dos usuários

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.

Antipadrões comuns:

- Seus desenvolvedores implantaram um novo recurso sem telemetria do usuário, e a utilização aumentou. Não é possível determinar se o aumento da utilização é proveniente do uso do novo recurso ou se é um problema introduzido com o novo código.
- Seus desenvolvedores implantaram um novo recurso sem telemetria do usuário. Não é possível saber se os clientes estão usando o recurso sem entrar em contato e perguntar a eles.

Benefícios do estabelecimento desta prática recomendada: Entenda como seus clientes usam seu aplicativo para identificar padrões de uso, comportamentos inesperados e permitir que você responda, se necessário.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Implantar a telemetria de atividades de usuários: use o código da aplicação para emitir informações sobre as atividades dos usuários (como cliques de streams 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.

OPS04-BP04 Implementar a telemetria de dependências

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.

Antipadrões comuns:

- Não é possível determinar se o motivo pelo qual seu aplicativo está inacessível é um problema de DNS sem executar manualmente uma verificação para ver se o provedor de DNS está funcionando.
- Seu aplicativo de carrinho de compras não consegue concluir transações. Não é possível determinar se há um problema com o provedor de processamento do seu cartão de crédito sem entrar em contato com ele para verificar.

Benefícios do estabelecimento desta prática recomendada: Entender a integridade das suas dependências permite que você responda, se necessário.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Implementar a telemetria de dependências: projete e configure a workload para emitir informações sobre o estado e o status dos sistemas dos quais depende. Alguns exemplos incluem: bancos de

dados externos, DNS, conectividade de rede e serviços externos de processamento de cartão de crédito.

- Amazon CloudWatch Agent with AWS Systems Manager integration - unified metrics & log collection for Linux & Windows (Integração do agente do Amazon CloudWatch com o AWS System Manager: métricas unificadas e coleta de logs para Linux e Windows)
- Collect metrics and logs from Amazon EC2 instances and on-premises servers with the CloudWatch Agent (Coletar métricas e logs das instâncias do Amazon EC2 e de servidores on-premises com o agente do CloudWatch)

Recursos

Documentos relacionados:

- Amazon CloudWatch Agent with AWS Systems Manager integration - unified metrics & log collection for Linux & Windows (Integração do agente do Amazon CloudWatch com o AWS System Manager: métricas unificadas e coleta de logs para Linux e Windows)
- Collect metrics and logs from Amazon EC2 instances and on-premises servers with the CloudWatch Agent (Coletar métricas e logs das instâncias do Amazon EC2 e de servidores on-premises com o agente do CloudWatch)

Exemplos relacionados:

- Laboratórios do Well-Architected: Monitoramento de dependências

OPS04-BP05 Implementar a capacidade de rastreamento das transações

Implemente o código do aplicativo e configure os componentes da carga de trabalho para emitir informações sobre o fluxo de transações na carga de trabalho. Use essas informações para determinar quando uma resposta é necessária e para identificar a causa raiz dos problemas.

Na AWS, é possível usar serviços de rastreamento distribuído, como o AWS X-Ray, para coletar e registrar rastreamentos à medida que as transações percorrem sua carga de trabalho, gerar mapas para ver como as transações fluem na carga de trabalho e serviços, obter informações sobre as relações entre componentes e identificar e analisar problemas em tempo real.

Antipadrões comuns:

- Você implementou uma arquitetura de microsserviços sem servidor que abrange várias contas. Seus clientes estão enfrentando problemas de performance intermitente. Você não consegue

descobrir qual função ou componente é responsável porque não há rastreamentos que permitiriam identificar onde no aplicativo está o problema de performance e o que está causando esse problema.

- Você está tentando determinar onde estão os gargalos de performance em sua carga de trabalho para que eles possam ser resolvidos em seus esforços de desenvolvimento. Não é possível ver a relação entre os componentes do aplicativo e os serviços com os quais eles interagem para determinar onde estão os gargalos, pois você não tem os rastreamentos que permitiriam analisar os serviços e caminhos específicos que afetam a performance do aplicativo.

Benefícios do estabelecimento desta prática recomendada: Entender o fluxo de transações em toda a carga de trabalho permite compreender o comportamento esperado das transações da carga de trabalho e as variações do comportamento esperado em toda a carga de trabalho, permitindo que você responda, se necessário.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Projete as aplicações e a workload para emitirem informações sobre o fluxo de transações entre os componentes do sistema, como o estágio da transação, o componente ativo e o tempo para concluir a atividade. Use essas informações para determinar o que está em andamento, o que está concluído e quais são os resultados das atividades concluídas. Isso ajuda a determinar quando uma resposta é necessária. Tempos de resposta da transação maiores que o esperado em um componente, por exemplo, podem indicar problemas com esse componente.
  - AWS X-Ray
  - O que é o AWS X-Ray?

Recursos

Documentos relacionados:

- AWS X-Ray
- O que é o AWS X-Ray?

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.

Práticas recomendadas

- OPS05-BP01 Usar o controle de versão
- OPS05-BP02 Testar e validar as alterações
- OPS05-BP03 Usar sistemas de gerenciamento de configuração
- OPS05-BP04 Usar sistemas de gerenciamento de compilação e de implantação
- OPS05-BP05 Executar o gerenciamento de patches
- OPS05-BP06 Compartilhar os padrões de design
- OPS05-BP07 Implementar práticas para aprimorar a qualidade do código
- OPS05-BP08 Usar vários ambientes
- OPS05-BP09 Fazer alterações frequentes, pequenas e reversíveis
- OPS05-BP10 Automatizar totalmente a integração e a implantação

OPS05-BP01 Usar o controle de versão

Use o controle de versão para habilitar o rastreamento de alterações e liberações.

Muitos serviços da AWS oferecem recursos de controle de versão. Use um sistema de revisão ou controle de origem como o AWS CodeCommit para gerenciar código e outros artefatos, como modelos do AWS CloudFormation com controle de versão da sua infraestrutura.

Antipadrões comuns:

- Você está desenvolvendo e armazenando seu código na estação de trabalho. Você teve uma falha de armazenamento irrecuperável na estação de trabalho em que seu código foi perdido.
- Depois de substituir o código existente pelas alterações, você reinicia o aplicativo e ele deixa de ser operável. Não é possível reverter para a alteração.
- Você tem um bloqueio de gravação em um arquivo de relatório que outra pessoa precisa editar. Ela entra em contato com você solicitando que você interrompa o trabalho para que ela possa concluir as tarefas.

- Sua equipe de pesquisa tem trabalhado em uma análise detalhada que moldará seu trabalho futuro. Alguém salvou acidentalmente sua lista de compras no relatório final. Não é possível reverter a alteração e você terá que recriar o relatório.

Benefícios do estabelecimento desta prática recomendada: Ao usar recursos de controle de versão, você pode reverter facilmente para bons estados conhecidos, versões anteriores e limitar o risco de perda de ativos.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Usar o controle de versão: mantenha os ativos em repositórios controlados por versão. Fazer isso oferece suporte para o rastreamento de alterações, a implantação de novas versões, a detecção de alterações nas versões existentes e a reversão para versões anteriores (por exemplo, a reversão para um estado bom e conhecido no caso de uma falha). Integre os recursos de controle de versão dos sistemas de gerenciamento de configurações aos seus procedimentos.
  - Introduction to AWS CodeCommit (Introdução ao AWS CodeCommit)
  - O que é o AWS CodeCommit?

Recursos

Documentos relacionados:

- O que é o AWS CodeCommit?

Vídeos relacionados:

- Introduction to AWS CodeCommit (Introdução ao AWS CodeCommit)

OPS05-BP02 Testar e validar as 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.

Muitos serviços da AWS oferecem recursos de controle de versão. Use um sistema de revisão ou controle de origem como o AWS CodeCommit para gerenciar código e outros artefatos, como modelos do AWS CloudFormation com controle de versão da sua infraestrutura.

Antipadrões comuns:

- Ao implantar novo código na produção, os clientes começam a ligar porque a aplicação não está mais funcionando.
- Você aplica novos grupos de segurança para aprimorar a segurança do perímetro. Isso funciona com consequências indesejadas. Os usuários não conseguem acessar as aplicações.
- Você modifica um método invocado pela nova função. Outra função também dependia desse método e não funciona mais. O problema não é detectado e entra em produção. A outra função não é invocada por algum tempo e, finalmente, falha na produção sem qualquer correlação com a causa.

Benefícios do estabelecimento desta prática recomendada: Ao testar e validar alterações antecipadamente, você pode resolver problemas com custos reduzidos e limitar o impacto sobre seus clientes. Ao testar antes da implantação, você reduz a possibilidade de erros.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Testar e validar as alterações: as alterações devem ser testadas, e os resultados validados, em todas as etapas do ciclo de vida (por exemplo, desenvolvimento, teste e produção). Use os resultados dos testes para confirmar novos recursos e reduzir o risco e o impacto de implantações com falha. Automatize os testes e a validação para garantir a consistência da análise, reduzir erros causados por processos manuais e reduzir o nível de esforço.
  - O que é o AWS CodeBuild?
  - Suporte de compilação local do AWS CodeBuild

Recursos

Documentos relacionados:

- Ferramentas do desenvolvedor da AWS
- Suporte de compilação local do AWS CodeBuild
- O que é o AWS CodeBuild?

OPS05-BP03 Usar sistemas de gerenciamento de configuração

Use os sistemas de gerenciamento de configuração 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.

O gerenciamento da configuração estática define valores ao inicializar um recurso que deve permanecer consistente durante todo o tempo de vida do recurso. Alguns exemplos incluem a definição da configuração do servidor web ou de aplicação em uma instância, ou a definição da configuração de um serviço da AWS no AWS Management Console ou por meio da AWS CLI.

O gerenciamento da configuração dinâmica define valores na inicialização que podem ou devem ser alterados durante o tempo de vida de um recurso. Por exemplo, é possível definir a alternância de um recurso para ativar uma funcionalidade no código por meio de uma alteração na configuração, ou alterar o nível de detalhes do log durante um incidente para capturar mais dados e alterar de volta depois do incidente, eliminando os logs agora desnecessários e a despesa associada.

Se tiver configurações dinâmicas em suas aplicações executadas em instâncias, contêineres, funções de tecnologia sem servidor ou dispositivos, você poderá usar o AWS AppConfig para gerenciar e implantá-las entre seus ambientes.

No AWS, você pode usar o AWS Config para monitorar continuamente as configurações de seus recursos da AWS entre contas e regiões. Isso permite rastrear o histórico da configuração, compreender como a alteração de uma configuração afeta outros recursos e auditá-la em relação a configurações esperadas ou desejadas, usando o Regras do AWS Config e os pacotes de conformidade do AWS Config.

Na AWS, é possível criar pipelines de integração contínua/implantação contínua (CI/CD) usando serviços como as Ferramentas do desenvolvedor da AWS (por exemplo, AWS CodeCommit, AWS CodeBuild, AWS CodePipeline, AWS CodeDeploye aos AWS CodeStar).

Tenha um calendário de alterações e monitore quando atividades ou eventos comerciais ou operacionais significativos que estão planejados podem ser afetados pela implementação da alteração. Ajuste as atividades para gerenciar riscos relacionados a esses planos. Calendário de alterações do AWS Systems Manager fornece um mecanismo para documentar blocos de tempo como abertos ou fechados para alterações e o motivo desses eventos, bem como para compartilhar essas informações com outras Contas da AWS. Os scripts do AWS Systems Manager Automation podem ser configurados para aderir ao estado de calendário de alteração.

AWS Systems Manager Maintenance Windows pode ser usado para programar a execução de scripts de Run Command ou de Automação do AWS SSM, invocações do AWS Lambda ou atividades do AWS Step Functions em horários especificados. Marque essas atividades no calendário de alterações para que elas possam ser incluídas na avaliação.

Antipadrões comuns:

- Você atualiza manualmente a configuração do servidor web em toda a frota e vários servidores não respondem devido a erros de atualização.
- Você atualiza manualmente a frota do servidor de aplicativos ao longo de muitas horas. A inconsistência na configuração durante a alteração causa comportamentos inesperados.
- Alguém atualizou seus grupos de segurança e seus servidores web não estão mais acessíveis. Sem saber o que foi alterado, você gasta muito tempo investigando o problema, ampliando o tempo de recuperação.

Benefícios do estabelecimento desta prática recomendada: A adoção de sistemas de gerenciamento de configurações reduz o nível de esforço para fazer e rastrear alterações, bem como a frequência de erros causados por procedimentos manuais.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Usar sistemas de gerenciamento de configuração: use sistemas de gerenciamento de configuração para rastrear e implementar alterações, reduzir erros causados por processos manuais e diminuir o nível de esforço.
  - Gerenciamento de configuração de infraestrutura
  - AWS Config
  - O que é o AWS Config?
  - Introduction to AWS CloudFormation (Introdução ao AWS CloudFormation)
  - O que é o AWS CloudFormation?
  - AWS OpsWorks
  - O que é o AWS OpsWorks?
  - Introduction to AWS Elastic Beanstalk (Introdução ao AWS Elastic Beanstalk)
  - O que é o AWS Elastic Beanstalk?

Recursos

Documentos relacionados:

- AWS AppConfig
- Ferramentas do desenvolvedor da AWS
- AWS OpsWorks
- Calendário de alterações do AWS Systems Manager
- AWS Systems Manager Maintenance Windows
- Gerenciamento de configuração de infraestrutura
- O que é o AWS CloudFormation?
- O que é o AWS Config?
- O que é o AWS Elastic Beanstalk?
- O que é o AWS OpsWorks?

Vídeos relacionados:

- Introduction to AWS CloudFormation (Introdução ao AWS CloudFormation)
- Introduction to AWS Elastic Beanstalk (Introdução ao AWS Elastic Beanstalk)

OPS05-BP04 Usar sistemas de gerenciamento de compilação e de 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.

Na AWS, é possível criar pipelines de integração contínua/implantação contínua (CI/CD) usando serviços como: Ferramentas do desenvolvedor da AWS (por exemplo, AWS CodeCommit, AWS CodeBuild, AWS CodePipeline, AWS CodeDeploye aos AWS CodeStar).

Antipadrões comuns:

- Depois de compilar o código no sistema de desenvolvimento e copiar o executável nos sistemas de produção, há uma falha na inicialização. Os arquivos de registro locais indicam que ele falhou devido à ausência de dependências.
- Você cria sua aplicação com êxito com os novos recursos em seu ambiente de desenvolvimento e fornece o código à garantia de qualidade (QA). Ele falha na QA porque não há ativos estáticos.

- Na sexta-feira, após muito esforço, você consegue criar o aplicativo manualmente em seu ambiente de desenvolvimento, incluindo os recursos recém-codificados. Na segunda-feira, você não consegue repetir as etapas que permitiram criar a aplicação com êxito.
- Você executa os testes que criou para a nova versão. Então você passa a próxima semana configurando um ambiente de teste e executando todos os testes de integração existentes, seguidos pelos testes de performance. O novo código tem um impacto inaceitável na performance e deve ser desenvolvido e testado novamente.

Benefícios do estabelecimento desta prática recomendada: Ao fornecer mecanismos para gerenciar atividades de criação e implantação, você reduz o nível de esforço para executar tarefas repetitivas, libera os membros da equipe para se concentrarem em tarefas criativas de alto valor e limita o surgimento de erros provenientes de procedimentos manuais.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Usar sistemas de gerenciamento de compilação e de implantação: use sistemas de gerenciamento de compilação e de implantação para rastrear e implementar alterações, reduzir erros causados por processos manuais e reduzir o nível de esforço. Automatize totalmente o pipeline de integração e implantação desde o check-in do código até a compilação, teste, implantação e validação. Isso reduz o tempo de execução, permite maior frequência de mudança e reduz o nível de esforço.
  - O que é o AWS CodeBuild?
  - As melhores práticas de integração contínua para equipes de desenvolvimento de software
  - Slalom: CI/CD para aplicações de tecnologia sem servidor na AWS
  - Introduction to AWS CodeDeploy: automated software deployment with Amazon Web Services (Introdução ao AWS CodeDeploy: implantação de software automatizada com a Amazon Web Services)
  - O que é o AWS CodeDeploy?

Recursos

Documentos relacionados:

- Ferramentas do desenvolvedor da AWS

- O que é o AWS CodeBuild?
- O que é o AWS CodeDeploy?

Vídeos relacionados:

- As melhores práticas de integração contínua para equipes de desenvolvimento de software
- Introduction to AWS CodeDeploy: automated software deployment with Amazon Web Services (Introdução ao AWS CodeDeploy: implantação de software automatizada com a Amazon Web Services)
- Slalom: CI/CD para aplicações de tecnologia sem servidor na AWS

OPS05-BP05 Executar o 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.

O gerenciamento de patches e vulnerabilidades faz parte de suas atividades de gerenciamento de benefícios e riscos. É preferível ter infraestruturas imutáveis e implantar cargas de trabalho em bons estados verificados e conhecidos. Quando isso não é viável, a aplicação de patches é a opção restante.

Atualizar imagens de máquinas, imagens de contêineres ou o Lambda tempos de execução personalizados e bibliotecas adicionais do Lambda para remover vulnerabilidades faz parte do gerenciamento de patches. Você deve gerenciar atualizações em Amazon Machine Images (AMIs) para imagens do Linux ou Windows Server usando o EC2 Image Builder. Você pode usar Amazon Elastic Container Registry com seu pipeline existente para gerenciar imagens do Amazon ECS e gerenciar imagens do Amazon EKS. O AWS Lambda inclui recursos de gerenciamento de versão .

A aplicação de patches não deve ser realizada em sistemas de produção sem antes testar em um ambiente seguro. Os patches só deverão ser aplicados se forem compatíveis com um resultado operacional ou comercial. No AWS, você pode usar o AWS Systems Manager Patch Manager para automatizar o processo de aplicação de patches em sistemas gerenciados e programar a atividade usando o AWS Systems Manager Maintenance Windows.

Antipadrões comuns:

- Você recebe uma ordem para aplicar todos os novos patches de segurança em até duas horas, resultando em várias interrupções devido à incompatibilidade da aplicação com os patches.
- Uma biblioteca sem patches resulta em consequências indesejadas, pois partes desconhecidas usam vulnerabilidades dentro dela para acessar sua carga de trabalho.
- Você aplica patches nos ambientes do desenvolvedor automaticamente, sem notificar os desenvolvedores. Você recebe várias reclamações dos desenvolvedores, dizendo que o ambiente deles deixa de funcionar conforme o esperado.
- Você não aplicou os patches no software pronto para uso comercial em uma instância persistente. Quando você tiver um problema com o software e entrar em contato com o fornecedor, ele informará que a versão não é compatível e será necessário aplicar patches a um nível específico para receber assistência.
- Um patch lançado recentemente para o software de criptografia que você usou tem melhorias significativas de performance. Seu sistema sem patches tem problemas de performance que permanecem enquanto não for feita a aplicação de patches.

Benefícios do estabelecimento desta prática recomendada: Ao estabelecer um processo de gerenciamento de patches, incluindo critérios de aplicação de patches e metodologia para distribuição em seus ambientes, você poderá perceber os benefícios e controlar o impacto. Isso permitirá a adoção de recursos e capacidades desejados, a remoção de problemas e a conformidade contínua com a governança. Implemente sistemas de gerenciamento de patches e automação para reduzir o nível de esforço para implantar patches e limitar erros causados por processos manuais.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Gerenciamento de patches: aplique patches aos sistemas para corrigir problemas, obter as capacidades ou os recursos desejados e permanecer em conformidade com a política de governança e com os requisitos de suporte do fornecedor. Em sistemas imutáveis, implante com o conjunto de patches adequado para alcançar o resultado desejado. Automatize o mecanismo de gerenciamento de patches para reduzir o tempo decorrido para aplicar patches, reduzir erros causados por processos manuais e reduzir o nível de esforço para corrigir.
  - AWS Systems Manager Patch Manager

Recursos

Documentos relacionados:

- Ferramentas do desenvolvedor da AWS
- AWS Systems Manager Patch Manager

Vídeos relacionados:

- CI/CD for Serverless Applications on AWS (CI/CD para aplicações de tecnologia sem servidor na AWS)
- Projeto com Ops em mente

Exemplos relacionados:

- Well-Architected Labs – Inventory and Patch Management (Laboratórios do Well-Architected: Gerenciamento de inventário e patches)

OPS05-BP06 Compartilhar os padrões de design

Compartilhe as melhores práticas entre as equipes para aumentar a conscientização e maximizar os benefícios dos esforços de desenvolvimento.

Na AWS, aplicativos, computação, infraestrutura e operações podem ser definidos e gerenciados usando metodologias de código. Isso permite fácil liberação, compartilhamento e adoção.

Muitos serviços e recursos da AWS foram projetados para serem compartilhados entre contas, permitindo que você compartilhe aprendizados e ativos criados com suas equipes. Por exemplo, você pode compartilhar repositórios do CodeCommit , funções do Lambda , buckets do Amazon S3e aos AMIs com contas específicas.

Ao publicar novos recursos ou atualizações, use o Amazon SNS para fornecer notificações entre contas. Os assinantes podem usar o Lambda para obter novas versões.

Se houver padrões compartilhados na sua organização, será fundamental a presença de mecanismos para solicitar adições, alterações e exceções para os padrões em suporte às atividades das equipes. Sem essa opção, os padrões se tornam uma restrição à inovação.

Antipadrões comuns:

- Você criou seu próprio mecanismo de autenticação de usuário, assim como cada uma das outras equipes de desenvolvimento em sua organização. Seus usuários precisam manter um conjunto separado de credenciais para cada parte do sistema que desejam acessar.

- Você criou seu próprio mecanismo de autenticação de usuário, assim como cada uma das outras equipes de desenvolvimento em sua organização. Sua organização recebe um novo requisito de conformidade que deve ser atendido. Agora, cada equipe de desenvolvimento deve investir os recursos para implementar o novo requisito.
- Você criou seu próprio layout de tela, assim como cada uma das outras equipes de desenvolvimento em sua organização. Seus usuários estão reclamando sobre a dificuldade de navegar pelas interfaces inconsistentes.

Benefícios do estabelecimento desta prática recomendada: Use padrões compartilhados para apoiar a adoção de melhores práticas e aumentar os benefícios dos esforços de desenvolvimento em que os padrões atendem aos requisitos de vários aplicativos ou organizações.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Compartilhar os padrões de design: compartilhe as práticas recomendadas existentes, os padrões de design, as listas de verificação, os procedimentos operacionais e os requisitos de orientações e de governança entre as equipes para reduzir a complexidade e maximizar os benefícios dos esforços de desenvolvimento. Garanta a existência de procedimentos para solicitar alterações, acréscimos e exceções aos padrões de projeto para apoiar a melhoria e inovação contínuas. Garanta que as equipes estejam cientes do conteúdo publicado, para que possam tirar proveito do conteúdo e limitar o retrabalho e o esforço desperdiçado.
  - Delegação de acesso ao ambiente da AWS
  - Compartilhar um repositório do AWS CodeCommit
  - Fácil autorização das funções do AWS Lambda
  - Compartilhamento de uma AMI com Contas da AWS específicas
  - Acelerar o compartilhamento de modelos com uma URL do designer do AWS CloudFormation
  - Usar o AWS Lambda com o Amazon SNS

Recursos

Documentos relacionados:

- Fácil autorização das funções do AWS Lambda
- Compartilhar um repositório do AWS CodeCommit

- Compartilhamento de uma AMI com Contas da AWS específicas
- Acelerar o compartilhamento de modelos com uma URL do designer do AWS CloudFormation
- Usar o AWS Lambda com o Amazon SNS

Vídeos relacionados:

- Delegação de acesso ao ambiente da AWS

OPS05-BP07 Implementar práticas para aprimorar a qualidade do código

Implemente práticas para aprimorar a qualidade do código e minimizar os defeitos. Alguns exemplos incluem desenvolvimento orientado por testes, análises de código e adoção de padrões.

Na AWS, é possível integrar serviços, como o Amazon CodeGuru , com o pipeline para identificar automaticamente os problemas potenciais de código e de segurança usando a análise de programa e o machine learning. O CodeGuru fornece orientações de como implementar as práticas recomendadas da AWS para resolver esses problemas.

Antipadrões comuns:

- Para poder testar seu recurso precocemente, você decidiu não integrar a biblioteca padrão de tratamento de entradas. Depois de testar, você confirma o código sem se lembrar de concluir a incorporação da biblioteca.
- Você tem pouca experiência com o conjunto de dados que está processando e não sabe que pode existir uma série de casos de borda no seu conjunto de dados. Esses casos de borda não são compatíveis com o código que você implementou.

Benefícios do estabelecimento desta prática recomendada: Com a adoção das práticas para melhorar a qualidade do código, é possível minimizar os problemas ocorridos na produção.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Implementar práticas para melhorar a qualidade do código: implemente práticas para melhorar a qualidade do código para minimizar os defeitos e o risco de serem implantados. Por exemplo, desenvolvimento orientado por testes, programação em pares, análises de código e adoção de padrões.

- Amazon CodeGuru

Recursos

Documentos relacionados:

- Amazon CodeGuru

OPS05-BP08 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.

Antipadrões comuns:

- Você está realizando o desenvolvimento em um ambiente de desenvolvimento compartilhado e outro desenvolvedor substitui suas alterações de código.
- Os controles de segurança restritivos em seu ambiente de desenvolvimento compartilhado estão impedindo que você experimente novos serviços e recursos.
- Você realiza testes de carga em seus sistemas de produção e causa uma interrupção para seus usuários.
- Ocorreu um erro crítico na produção que resulta na perda de dados. No ambiente de produção, você tenta recriar as condições que levaram à perda de dados para identificar como isso aconteceu e impedir a recorrência. Para evitar mais perda de dados durante o teste, você é forçado a tornar a aplicação indisponível para os usuários.
- Você está operando um serviço multilocatário e não consegue oferecer suporte a uma solicitação do cliente para um ambiente dedicado.
- Talvez você não teste sempre, mas quando o faz, já está em produção.
- Você acredita que a simplicidade de um único ambiente substitui o escopo do impacto das alterações dentro do ambiente.

Benefícios do estabelecimento desta prática recomendada: Ao implantar vários ambientes, você pode oferecer suporte a vários ambientes simultâneos de desenvolvimento, teste e produção, sem criar conflitos entre desenvolvedores ou comunidades de usuários.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Usar vários ambientes: forneça aos desenvolvedores ambientes de área restrita para testes com controles minimizados para permitir a experimentação. Forneça ambientes de desenvolvimento individuais para permitir o trabalho em paralelo, aumentando a agilidade do desenvolvimento. Implemente controles mais rigorosos nos ambientes ao se aproximar da produção para permitir que os desenvolvedores inovem. Use a infraestrutura como sistemas de gerenciamento de código e configuração para implantar ambientes que são configurados de maneira consistente com os controles presentes na produção para garantir que os sistemas operem conforme o esperado quando implantados. Quando os ambientes não estiverem em uso, desligue-os para evitar custos associados a recursos inativos (por exemplo, sistemas de desenvolvimento à noite e fins de semana). Implante ambientes equivalentes de produção ao carregar o teste para habilitar resultados válidos.
  - O que é o AWS CloudFormation?
  - Como interrompo e inicio instâncias do Amazon EC2 em intervalos regulares usando o AWS Lambda?

Recursos

Documentos relacionados:

- Como interrompo e inicio instâncias do Amazon EC2 em intervalos regulares usando o AWS Lambda?
- O que é o AWS CloudFormation?

OPS05-BP09 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.

Antipadrões comuns:

- Você implanta uma nova versão do seu aplicativo trimestralmente.
- Você faz alterações no esquema de banco de dados com frequência.
- Você realiza atualizações manuais no local, substituindo instalações e configurações existentes.

Benefícios do estabelecimento desta prática recomendada: Você reconhece os benefícios dos esforços de desenvolvimento mais rapidamente implantando pequenas alterações com frequência. Quando as alterações são pequenas, é muito mais fácil identificar se elas têm consequências indesejadas. Quando as alterações são reversíveis, há menos risco de implementar a alteração à medida que a recuperação é simplificada.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- 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. Também aumenta a taxa na qual você pode agregar valor aos negócios.

OPS05-BP10 Automatizar 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.

Aplique metadados usando o Tags de recursos e AWS Resource Groups seguindo uma estratégia de marcação consistente para permitir a identificação dos seus recursos. Identifique seus recursos para organização, contabilidade de custos, controles de acesso e direcione a execução de atividades operacionais automatizadas.

Antipadrões comuns:

- Na sexta-feira, você conclui a criação do novo código para a ramificação do recurso. Na segunda-feira, depois de executar os scripts de teste de qualidade em cada um dos scripts de testes unitários, você verificará o código para o próximo lançamento programado.
- Você tem a tarefa de codificar uma correção para um problema crítico que afeta um grande número de clientes em produção. Depois de testar a correção, você confirma o gerenciamento de alterações de e-mail e do código para solicitar aprovação para implantação na produção.

Benefícios do estabelecimento desta prática recomendada: Ao implementar sistemas automatizados de gerenciamento de criação e implantação, você reduz os erros causados por processos manuais e o esforço para implantar alterações, permitindo que os membros da equipe se concentrem na entrega de valor empresarial.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Usar sistemas de gerenciamento de compilação e de implantação: use sistemas de gerenciamento de compilação e de implantação para rastrear e implementar alterações, reduzir erros causados por processos manuais e reduzir o nível de esforço. Automatize totalmente o pipeline de integração e implantação desde o check-in do código até a compilação, teste, implantação e validação. Isso reduz o tempo de execução, permite maior frequência de mudança e reduz o nível de esforço.
  - O que é o AWS CodeBuild?
  - As melhores práticas de integração contínua para equipes de desenvolvimento de software
  - Slalom: CI/CD para aplicações de tecnologia sem servidor na AWS
  - Introduction to AWS CodeDeploy: automated software deployment with Amazon Web Services (Introdução ao AWS CodeDeploy: implantação de software automatizada com a Amazon Web Services)
  - O que é o AWS CodeDeploy?

Recursos

Documentos relacionados:

- O que é o AWS CodeBuild?
- O que é o AWS CodeDeploy?

Vídeos relacionados:

- As melhores práticas de integração contínua para equipes de desenvolvimento de software
- Introduction to AWS CodeDeploy: automated software deployment with Amazon Web Services (Introdução ao AWS CodeDeploy: implantação de software automatizada com a Amazon Web Services)
- Slalom: CI/CD para aplicações de tecnologia sem servidor na AWS

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.

Práticas recomendadas

- OPS06-BP01 Planejar para alterações malsucedidas
- OPS06-BP02 Testar e validar as alterações
- OPS06-BP03 Usar sistemas de gerenciamento para implantação
- OPS06-BP04 Testar usando implantações limitadas
- OPS06-BP05 Implantar usando ambientes paralelos
- OPS06-BP06 Implantar alterações frequentes, pequenas e reversíveis
- OPS06-BP07 Automatizar totalmente a integração e a implantação
- OPS06-BP08 Automatizar os testes e a reversão

OPS06-BP01 Planejar para alterações malsucedidas

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.

Antipadrões comuns:

- Você executou uma implantação e seu aplicativo se tornou instável, mas parece haver usuários ativos no sistema. Você precisa decidir se deseja reverter a alteração e afetar os usuários ativos ou esperar para reverter a alteração sabendo que mesmo assim os usuários podem ser afetados.
- Depois de fazer uma alteração de rotina, os novos ambientes ficam acessíveis, mas uma de suas sub-redes se tornou inacessível. Você precisa decidir se deseja reverter tudo ou tentar corrigir a sub-rede inacessível. Enquanto você estiver fazendo essa determinação, a sub-rede permanece inacessível.

Benefícios do estabelecimento desta prática recomendada: Quando há um plano estabelecido para reduzir o tempo médio de recuperação (MTTR) de alterações malsucedidas, minimizando o impacto para os usuários finais.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Planejar para alterações malsucedidas: planeje para reverter para um bom estado conhecido (ou seja, reverter a alteração) ou realizar reparos no ambiente de produção (ou seja, avançar com a alteração) se uma alteração não tiver o resultado desejado. Ao identificar alterações que não podem ser revertidas se mal-sucedidas, aplique a auditoria devida antes de confirmar a alteração.

OPS06-BP02 Testar e validar as 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.

Na AWS, você pode criar ambientes paralelos temporários para reduzir o risco, o esforço e o custo da experimentação e dos testes. Automatize a implantação desses ambientes usando o AWS CloudFormation para garantir implementações consistentes dos seus ambientes temporários.

Antipadrões comuns:

- Você implanta um novo recurso incrível em seu aplicativo. Ele não funciona. Você não sabe.
- Você atualiza seus certificados. Você instala acidentalmente os certificados nos componentes incorretos. Você não sabe.

Benefícios do estabelecimento desta prática recomendada: Ao testar e validar as alterações após a implantação, você pode identificar os problemas antecipadamente, oferecendo a oportunidade de reduzir o impacto sobre seus clientes.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Testar e validar as alterações: teste as alterações e valide os resultados em todas as etapas do ciclo de vida (como desenvolvimento, teste e produção) a fim de confirmar novos recursos e minimizar o risco e o impacto de implantações com falha.
  - AWS Cloud9
  - O que é o AWS Cloud9?
  - Como testar e depurar o AWS CodeDeploy localmente antes de enviar o código

Recursos

Documentos relacionados:

- AWS Cloud9
- Ferramentas do desenvolvedor da AWS
- Como testar e depurar o AWS CodeDeploy localmente antes de enviar o código
- O que é o AWS Cloud9?

OPS06-BP03 Usar 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.

Na AWS, é possível criar pipelines de integração contínua/implantação contínua (CI/CD) usando serviços como: Ferramentas do desenvolvedor da AWS (por exemplo, AWS CodeCommit, AWS CodeBuild, AWS CodePipeline, AWS CodeDeploye aos AWS CodeStar).

Antipadrões comuns:

- Você implanta atualizações manualmente nos servidores de aplicativos em toda a frota e vários servidores não respondem devido a erros de atualização.
- Você implanta manualmente a frota do servidor de aplicativos ao longo de muitas horas. A inconsistência nas versões durante a alteração causa comportamentos inesperados.

Benefícios do estabelecimento desta prática recomendada: A adoção de sistemas de gerenciamento de implantação reduz o nível de esforço para implantar alterações e a frequência de erros causados por procedimentos manuais.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Usar sistemas de gerenciamento de implantação: use sistemas de gerenciamento de implantação para monitorar e implementar alterações. Isso reduzirá os erros causados pelos processos manuais e o nível de esforço para implantar as alterações. Automatize o pipeline de integração e implantação desde o check-in do código até o teste, a implantação e a validação. Isso reduz o tempo de execução, permite maior frequência de mudança e reduz ainda mais o nível de esforço.

- Introduction to AWS CodeDeploy: automated software deployment with Amazon Web Services (Introdução ao AWS CodeDeploy: implantação de software automatizada com a Amazon Web Services)
- O que é o AWS CodeDeploy?
- O que é o AWS Elastic Beanstalk?
- O que é o Amazon API Gateway?

Recursos

Documentos relacionados:

- Guia do usuário do AWS CodeDeploy
- Ferramentas do desenvolvedor da AWS
- Experimentar uma amostra da implantação azul/verde no AWS CodeDeploy
- O que é o AWS CodeDeploy?
- O que é o AWS Elastic Beanstalk?
- O que é o Amazon API Gateway?

Vídeos relacionados:

- Deep Dive on Advanced Continuous Delivery Techniques Using AWS (Mergulhe nas técnicas avançadas de entrega contínua usando a AWS)
- Introduction to AWS CodeDeploy: automated software deployment with Amazon Web Services (Introdução ao AWS CodeDeploy: implantação de software automatizada com a Amazon Web Services)

OPS06-BP04 Testar 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.

Antipadrões comuns:

- Você implanta uma alteração malsucedida em toda a produção de uma só vez. Você não sabe.

Benefícios do estabelecimento desta prática recomendada: Ao testar e validar as alterações após a implantação limitada, você pode identificar problemas antecipadamente com impacto mínimo em seus clientes, oferecendo a oportunidade de reduzir ainda mais o impacto sobre seus clientes.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Testar usando implantações limitadas: teste com implantações limitadas junto com sistemas existentes para confirmar os resultados desejados antes da implantação em grande escala. Use testes para implantação canário ou implantações individuais, por exemplo.
  - Guia do usuário do AWS CodeDeploy
  - Implantações azul/verde com o AWS Elastic Beanstalk
  - Configurar uma implantação de lançamento canário com o API Gateway
  - Experimentar uma amostra da implantação azul/verde no AWS CodeDeploy
  - Como trabalhar com configurações de implantação no AWS CodeDeploy

Recursos

Documentos relacionados:

- Guia do usuário do AWS CodeDeploy
- Implantações azul/verde com o AWS Elastic Beanstalk
- Configurar uma implantação de lançamento canário com o API Gateway
- Experimentar uma amostra da implantação azul/verde no AWS CodeDeploy
- Como trabalhar com configurações de implantação no AWS CodeDeploy

OPS06-BP05 Implantar 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.

Antipadrões comuns:

- Você executa uma implantação mutável modificando os sistemas existentes. Ao descobrir que a alteração não foi bem-sucedida, você será forçado a modificar os sistemas novamente para restaurar a versão antiga, aumentando o tempo de recuperação.
- Durante uma janela de manutenção, você desativará o ambiente antigo e começará a criar o novo ambiente. Muitas horas após o procedimento, você descobre problemas irrecuperáveis com a implantação. Embora extremamente cansado, você é forçado a encontrar os procedimentos de implantação anteriores e começar a reconstruir o ambiente antigo.

Benefícios do estabelecimento desta prática recomendada: Ao usar ambientes paralelos, é possível pré-implantar o novo ambiente e fazer a transição para ele quando desejar. Se o novo ambiente não for bem-sucedido, você poderá se recuperar rapidamente fazendo a transição de volta para o ambiente original.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Implantar 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. Isso minimiza o tempo de recuperação, permitindo assim a reversão para o ambiente anterior. Use infraestruturas imutáveis com implantações azul/verde, por exemplo.
  - Como trabalhar com configurações de implantação no AWS CodeDeploy
  - Implantações azul/verde com o AWS Elastic Beanstalk
  - Configurar uma implantação de lançamento canário com o API Gateway
  - Experimentar uma amostra da implantação azul/verde no AWS CodeDeploy

Recursos

Documentos relacionados:

- Guia do usuário do AWS CodeDeploy
- Implantações azul/verde com o AWS Elastic Beanstalk
- Configurar uma implantação de lançamento canário com o API Gateway
- Experimentar uma amostra da implantação azul/verde no AWS CodeDeploy
- Como trabalhar com configurações de implantação no AWS CodeDeploy

Vídeos relacionados:

- Deep Dive on Advanced Continuous Delivery Techniques Using AWS (Mergulhe nas técnicas avançadas de entrega contínua usando a AWS)

OPS06-BP06 Implantar alterações 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.

Antipadrões comuns:

- Você implanta uma nova versão do seu aplicativo trimestralmente.
- Você faz alterações no esquema de banco de dados com frequência.
- Você realiza atualizações manuais no local, substituindo instalações e configurações existentes.

Benefícios do estabelecimento desta prática recomendada: Você reconhece os benefícios dos esforços de desenvolvimento mais rapidamente implantando pequenas alterações com frequência. Quando as alterações são pequenas, é muito mais fácil identificar se elas têm consequências indesejadas. Quando as alterações são reversíveis, há menos risco de implementar a alteração à medida que a recuperação é simplificada.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Implantar alterações 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.

OPS06-BP07 Automatizar 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.

Aplique metadados usando o Tags de recursos e AWS Resource Groups seguindo uma estratégia de marcação consistente para permitir a identificação dos seus recursos. Identifique seus recursos

para organização, contabilidade de custos, controles de acesso e direcione a execução de atividades operacionais automatizadas.

Antipadrões comuns:

- Na sexta-feira, você conclui a criação do novo código para a ramificação do recurso. Na segunda-feira, depois de executar os scripts de teste de qualidade em cada um dos scripts de testes unitários, você verificará o código para o próximo lançamento programado.
- Você tem a tarefa de codificar uma correção para um problema crítico que afeta um grande número de clientes em produção. Depois de testar a correção, você confirma o gerenciamento de alterações de e-mail e do código para solicitar aprovação para implantação na produção.

Benefícios do estabelecimento desta prática recomendada: Ao implementar sistemas automatizados de gerenciamento de criação e implantação, você reduz os erros causados por processos manuais e o esforço para implantar alterações, permitindo que os membros da equipe se concentrem na entrega de valor empresarial.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Usar sistemas de gerenciamento de compilação e de implantação: use sistemas de gerenciamento de compilação e de implantação para rastrear e implementar alterações, reduzir erros causados por processos manuais e reduzir o nível de esforço. Automatize totalmente o pipeline de integração e implantação desde o check-in do código até a compilação, teste, implantação e validação. Isso reduz o tempo de execução, permite maior frequência de mudança e reduz o nível de esforço.
  - O que é o AWS CodeBuild?
  - As melhores práticas de integração contínua para equipes de desenvolvimento de software
  - Slalom: CI/CD para aplicações de tecnologia sem servidor na AWS
  - Introduction to AWS CodeDeploy: automated software deployment with Amazon Web Services (Introdução ao AWS CodeDeploy: implantação de software automatizada com a Amazon Web Services)
  - O que é o AWS CodeDeploy?
  - Deep Dive on Advanced Continuous Delivery Techniques Using AWS (Mergulhe nas técnicas avançadas de entrega contínua usando a AWS)

Recursos

Documentos relacionados:

- Experimentar uma amostra da implantação azul/verde no AWS CodeDeploy
- O que é o AWS CodeBuild?
- O que é o AWS CodeDeploy?

Vídeos relacionados:

- As melhores práticas de integração contínua para equipes de desenvolvimento de software
- Deep Dive on Advanced Continuous Delivery Techniques Using AWS (Mergulhe nas técnicas avançadas de entrega contínua usando a AWS)
- Introduction to AWS CodeDeploy: automated software deployment with Amazon Web Services (Introdução ao AWS CodeDeploy: implantação de software automatizada com a Amazon Web Services)
- Slalom: CI/CD para aplicações de tecnologia sem servidor na AWS

OPS06-BP08 Automatizar os testes e a reversão

Automatize os testes dos ambientes implantados para confirmar os resultados desejados. Automatize a reversão para um bom estado anterior conhecido quando os resultados não forem alcançados, para minimizar o tempo de recuperação e reduzir os erros causados por processos manuais.

Antipadrões comuns:

- Você implanta alterações em sua carga de trabalho. Depois de verificar se a alteração foi concluída, você inicia os testes de pós-implantação. Depois de concluídos, você percebe que sua workload está inoperante e que os clientes estão desconectados. Em seguida, você começa a reverter para a versão anterior. Depois de um período prolongado para detectar o problema, o tempo de recuperação é estendido pela reimplantação manual.

Benefícios do estabelecimento desta prática recomendada: Ao testar e validar alterações após a implantação, é possível identificar problemas imediatamente. Ao reverter automaticamente para a versão anterior, o impacto sobre os clientes é minimizado.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Automatizar testes e reversão: automatize testes de ambientes implantados para confirmar os resultados desejados. Automatize a reversão para um bom estado anterior conhecido quando os resultados não forem alcançados, para minimizar o tempo de recuperação e reduzir os erros causados por processos manuais. Por exemplo, faça transações sintéticas e detalhadas do usuário após a implantação, verifique os resultados e reverta a falha.
  - Reimplantar e reverter uma implantação com o AWS CodeDeploy

Recursos

Documentos relacionados:

- Reimplantar e reverter uma implantação com o AWS CodeDeploy

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.

Práticas recomendadas

- OPS07-BP01 Garantir a capacidade da equipe
- OPS07-BP02 Garantir uma análise consistente da prontidão operacional
- OPS07-BP03 Usar runbooks para realizar procedimentos
- OPS07-BP04 Usar manuais para investigar problemas
- OPS07-BP05 Tomar decisões embasadas para implantar sistemas e alterações

OPS07-BP01 Garantir a capacidade da equipe

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.

Você precisará ter membros da equipe suficientes para cobrir todas as atividades (inclusive em plantão). Garanta que suas equipes tenham as habilidades necessárias para terem êxito no treinamento sobre as workload, as ferramentas das operações e a AWS.

A AWS fornece recursos, incluindo o Centro de recursos de conceitos básicos da AWS, Blogs da AWS, AWS Online Tech Talks, Eventos e webinars da AWSe os Laboratórios do AWS Well-Architected, que fornecem orientações, exemplos e demonstrações detalhadas para educar suas equipes. Além disso, o Treinamento da AWS and Certification fornece algum treinamento gratuito por meio de cursos digitais autoguiados sobre os conceitos básicos da AWS. Também é possível inscrever-se em treinamento administrado por instrutor para oferecer suporte adicional ao desenvolvimento das habilidades em AWS de suas equipes.

Antipadrões comuns:

- Implantar uma carga de trabalho sem membros qualificados na equipe para oferecer suporte à plataforma e aos serviços em uso.
- Implantar uma carga de trabalho sem membros da equipe disponíveis durante as horas pretendidas de suporte.
- Implantar uma carga de trabalho sem membros suficientes da equipe para oferecer suporte se houver membros da equipe em licença ou afastados por doença.
- Implantar cargas de trabalho adicionais sem analisar o impacto adicional sobre os membros da equipe que oferecem suporte e outras cargas de trabalho.

Benefícios do estabelecimento desta prática recomendada: Ter membros da equipe qualificados possibilita o suporte eficaz da sua carga de trabalho.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Capacidade da equipe: valide se a equipe com treinamento é grande o suficiente para oferecer suporte de forma eficaz à workload.
  - Tamanho da equipe: verifique se você tem membros da equipe suficientes para cobrir as atividades operacionais, como tarefas de plantão.
  - Habilidades da equipe: verifique se os membros da equipe têm treinamento suficiente da AWS, de workload e de ferramentas operacionais para realizarem suas tarefas.
    - Eventos e webinars da AWS

    - Nossas boas-vindas ao Treinamento da AWS and Certification
- Analisar os recursos: analise o tamanho e as habilidades da equipe conforme as condições operacionais e as workloads mudam, para garantir que haja capacidade suficiente para manter a excelência operacional. Faça ajustes para garantir que o tamanho e a habilidade da equipe correspondam aos requisitos operacionais para as cargas de trabalho para as quais a equipe fornece suporte.

Recursos

Documentos relacionados:

- Blogs da AWS
- Eventos e webinars da AWS
- Centro de recursos de conceitos básicos da AWS
- AWS Online Tech Talks
- Nossas boas-vindas ao Treinamento da AWS and Certification

Exemplos relacionados:

- Laboratórios do Well-Architected

OPS07-BP02 Garantir uma análise consistente da prontidão operacional

Use Análises de prontidão operacional (ORRs) para validar que você pode operar sua workload. A ORR é um mecanismo desenvolvido na Amazon para validar que as equipes podem operar as workloads com segurança. Uma ORR é um processo de análise e inspeção que usa uma lista de verificação de requisitos. Uma ORR é uma experiência de autoatendimento que as equipes usam para certificar suas workloads. As ORRs incluem práticas recomendadas de lições aprendidas de nossos anos de experiência na criação de software.

Uma lista de verificação de ORR é composta de recomendações de arquitetura, processo operacional, gerenciamento de evento e qualidade de lançamento. Nosso processo de Correção de erros (CoE) é um motivador principal desses itens. Sua própria análise pós-incidente deve impulsionar a evolução de sua própria ORR. Uma ORR não é apenas sobre seguir as práticas recomendadas, mas evitar a recorrência de eventos que você já viu. Por fim, os requisitos de segurança, governança e conformidade também podem ser incluídos em uma ORR.

Execute ORRs antes do lançamento de uma workload para disponibilidade geral e por todo o ciclo de vida de desenvolvimento do software. A execução da ORR antes do lançamento aumenta a capacidade de operar a workload com segurança. Execute a ORR periodicamente na workload para identificar qualquer desvio das práticas recomendadas. Você pode ter listas de verificação da ORR para o lançamento de outros serviços e ORRs para avaliações periódicas. Isso ajuda a manter você atualizado sobre as novas práticas recomendadas que surgem e incorporar as lições aprendidas da análise pós-incidente. À medida que seu uso da nuvem amadurece, é possível criar requisitos de ORR em sua arquitetura como padrões.

Resultado desejado: você tem uma lista de verificação da ORR com as práticas recomendadas para sua organização. As ORRs são realizadas antes do lançamento das workloads. As ORRs são executadas periodicamente ao longo do ciclo de vida da workload.

Antipadrões comuns:

- Você lança uma workload sem saber se pode operá-la.
- Os requisitos de governança e segurança não estão incluídos na certificação de uma workload para o lançamento.
- As workloads não são reavaliadas periodicamente.
- As workloads são lançadas sem a aplicação dos procedimentos exigidos.
- Você vê a repetição das mesmas falhas da causa raiz em várias workloads.

Benefícios de estabelecer esta prática recomendada:

- suas workloads incluem práticas recomendadas de arquitetura, processo e gerenciamento.
- As lições aprendidas são incorporadas em seu processo de ORR.
- Os procedimentos exigidos estão em vigor no lançamento das workloads.
- As ORRs são executadas durante todo o ciclo de vida do software das workloads.

Nível de risco caso essa prática recomendada não seja estabelecida: alto

Orientação para implementação

Uma ORR é composta por dois elementos: um processo e uma lista de verificação. O processo da ORR deve ser adotado pela organização e ter o apoio de um patrocinador executivo. No mínimo, as ORRs devem ser realizadas antes do lançamento da workload para disponibilidade geral. Execute a

ORR ao longo de todo o ciclo de vida de desenvolvimento do software para mantê-la atualizada com as práticas recomendadas ou os novos requisitos. A lista de verificação da ORR deve incluir itens de configuração, requisitos de segurança e governança e práticas recomendadas de sua organização. Ao longo do tempo, você pode usar serviços como o AWS Config, o AWS Security Hube o AWS Control Tower Guardrails, para criar práticas recomendadas com base na ORR visando as barreiras de proteção para detecção automáticas das práticas recomendadas.

Exemplo de cliente

Depois de vários incidentes na produção, a Loja UmaEmpresa decidiu implementar um processo de ORR. Ela criou uma lista de verificação composta de práticas recomendadas, requisitos de governança e conformidade e lições aprendidas de interrupções. Novas workloads passam pelo processo de ORR antes do lançamento. É realizada uma ORR anualmente para cada workload com um subconjunto de práticas recomendadas a incorporar novas práticas recomendadas e requisitos que são adicionados à lista de verificação da ORR. Ao longo do tempo, a Loja UmaEmpresa usou o AWS Config para detectar algumas práticas recomendadas, acelerando o processo de ORR.

Etapas da implementação

Para saber mais sobre as ORRs, leia o whitepaper de Análises de prontidão operacional (ORR). Ele fornece informações detalhadas sobre o histórico do processo de ORR, como criar sua própria prática de ORR e como desenvolver sua lista de verificação da ORR. As etapas a seguir são uma versão resumida desse documento. Para uma compreensão aprofundada do que são as ORRs e de como criar sua própria, recomendamos a leitura desse whitepaper.

1. Reúna as principais partes interessadas, incluindo os representantes de segurança, operações e desenvolvimento.
2. Peça para cada parte interessada fornecer pelo menos um requisito. Para a primeira iteração, tente limitar o número de itens para trinta ou menos.
    - Apêndice B: os exemplos de perguntas da ORR do whitepaper de Análises de prontidão operacional (ORR) contém exemplos de perguntas que você pode usar para começar.
3. Reúna seus requisitos em uma planilha.
    - Você pode usar o Custom Lenses no AWS Well-Architected Tool para desenvolver sua ORR e compartilhá-la em suas contas e no AWS Organization.
4. Identifique uma workload na qual realizar a ORR. O ideal seria em uma workload em pré-lançamento ou uma workload interna.

5. Execute a lista de verificação completa da ORR e anote as descobertas feitas. As descobertas podem não ser corretas caso esteja ocorrendo uma mitigação. Para descobertas que não tenham uma mitigação, acrescente-as à sua lista de pendências e implemente-as antes do lançamento.
6. Continue a adicionar práticas recomendadas e requisitos à sua lista de verificação de ORR ao longo do tempo.

Os clientes do AWS Support com Enterprise Support podem solicitar o workshop de Análises de prontidão operacional com seu gerente de conta técnico. O workshop é uma sessão interativa de trabalho em retrospecto para que você consiga desenvolver sua própria lista de verificação de ORR.

Nível de esforço do plano de implementação: alto. Adotar uma prática de ORR em sua organização exige a adesão de um patrocinador executivo e das partes interessadas. Crie e atualize a lista de verificação com as opiniões de toda a sua organização.

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP03 Avaliar os requisitos de governança – Os requisitos de governança são uma opção natural para uma lista de verificação da ORR.
- OPS01-BP04 Avaliar os requisitos de conformidade – Os requisitos de conformidade, às vezes são incluídos em uma lista de verificação de ORR. Outras vezes, eles constituem um processo separado.
- OPS03-BP07 Fornecer recursos adequados às equipes – A capacidade da equipe é uma boa candidata para um requisito de ORR.
- OPS06-BP01 Planejar para alterações malsucedidas – Um plano de reversão ou avanço deve ser estabelecido antes do lançamento da workload.
- OPS07-BP01 Garantir a capacidade da equipe – Para comportar uma workload, você deve ter o pessoal necessário.
- SEC01-BP03 Identificar e validar objetivos de controle – Os objetivos de controle de segurança compõem excelentes requisitos de ORR.
- REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados – Os planos de recuperação de desastres são um ótimo requisito de ORR.
- COST02-BP01 Desenvolver políticas com base nos requisitos da sua organização – As políticas de gerenciamento de custos são ótimas para incluir em sua lista de verificação de ORR.

Documentos relacionados:

- AWS Control Tower - Guardrails in AWS Control Tower (AWS Control Tower: barreiras de proteção no AWS Control Tower)
- AWS Well-Architected Tool - Custom Lenses
- Operational Readiness Review Template by Adrian Hornsby (Modelo de Análise de prontidão operacional, por Adrian Hornsby)
- Whitepaper de Análises de prontidão operacional (ORR)

Vídeos relacionados:

- AWS Supports You | Building an Effective Operational Readiness Review (ORR) (Apoio do AWS Support: criação de uma Análise de prontidão operacional (ORR) eficaz)

Exemplos relacionados:

- Sample Operational Readiness Review (ORR) Lens (Exemplo da perspectiva da Análise de prontidão operacional (ORR))

Serviços relacionados:

- AWS Config
- AWS Control Tower
- AWS Security Hub
- AWS Well-Architected Tool

OPS07-BP03 Usar runbooks para realizar procedimentos

A runbook é um processo documentado para alcançar um resultado específico. Runbooks consistem em uma série de etapas que alguém segue para realizar alguma coisa. Runbooks são usados em operações desde os primórdios da aviação. Nas operações na nuvem, usamos runbooks para reduzir o risco e alcançar os resultados desejados. Em essência, um runbook é uma lista de verificação para concluir uma tarefa.

Runbooks são fundamentais para a operação de uma workload. Da integração de um novo membro da equipe à implantação de um lançamento importante, os runbooks são os processos codificados

que fornecem resultados consistentes independentemente de que os usa. Os runbooks devem estar publicados em um local central e devem ser atualizados à medida que o processo evolui, uma vez que a atualização dos runbooks é um aspecto fundamental de um processo de gerenciamento de mudanças. Também devem incluir orientação sobre tratamento de erros, ferramentas, permissões, exceções e encaminhamentos em caso de problema.

À medida que sua organização amadurece, comece a automatizar os runbooks. Comece com runbooks que sejam curtos e usados com frequência. Use linguagens de scripts para automatizar as etapas ou facilitar a realização delas. À medida que você automatiza os primeiros runbooks, vai dedicar tempo à automação de runbooks mais complexos. Com o tempo, a maioria dos seus runbooks deverão ter algum nível de automação.

Resultado desejado: sua equipe tem um conjunto de guias detalhados para realizar tarefas de workload. Os runbooks contêm o resultado desejado, as ferramentas e permissões necessárias e as instruções para tratamento de erros. Eles estão armazenados em um local central e são atualizados frequentemente.

Antipadrões comuns:

- Depender da memória para concluir cada etapa de um processo.
- Implantar mudanças manualmente sem uma lista de verificação.
- Vários membros da equipe realizando o mesmo processo, mas com etapas ou resultados diferentes.
- Deixar que os runbooks fiquem desatualizados em relação às mudanças no sistema e à automação.

Benefícios do estabelecimento desta prática recomendada:

- Redução das taxas de erros em tarefas manuais.
- Operações realizadas de maneira consistente.
- Novos membros da equipe podem começar a realizar tarefas mais cedo.
- Os runbooks podem ser automatizados para reduzir o esforço.

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

Orientação de implementação

Os runbooks podem assumir diversos formatos dependendo do nível de maturidade da sua organização. No mínimo, devem consistir em um documento de texto detalhado. O resultado desejado deve estar claramente identificado. Documentar claramente as permissões ou ferramentas especiais necessárias. Fornecer orientação detalhada sobre tratamento de erros e encaminhamentos em caso de problema. Listar o proprietário do runbook e publicá-lo em um local central. Depois que o runbook estiver documentado, valide-o pedindo que outro membro da equipe o execute. À medida que os procedimentos evoluem, atualize os runbooks de acordo com seu processo de gerenciamento de mudanças.

Os runbooks em texto devem ser automatizados à medida que a organização amadurece. Usando serviços como as automações do AWS Systems Manager, você pode transformar texto plano em automações que podem ser executadas na workload. Essas automações podem ser executadas em resposta a eventos, reduzindo a sobrecarga operacional de manutenção da workload.

Exemplo de cliente

A AnyCompany Retail precisa realizar atualizações no esquema de banco de dados durante implantações de software. A equipe de operações na nuvem trabalhou com a equipe de administração do banco de dados para criar um runbook para implantação manual dessas mudanças. O runbook lista cada etapa do processo em um formato de lista de verificação. Ele inclui uma seção sobre tratamento de erros em caso de problema. Eles publicaram o runbook na wiki interna junto com outros runbooks. A equipe de operações na nuvem planeja automatizar o runbook em um sprint futuro.

Etapas da implementação

Se você não tem um repositório de documentos, um repositório de controle de versão é um ótimo lugar para começar a criar sua biblioteca de runbooks. Você pode criar runbooks usando Markdown. Disponibilizamos um modelo de runbook que você pode usar para começar a criar runbooks.

```
# Título do runbook ## Informações do runbook | ID do runbook | Descrição | Ferramentas
 usadas | Permissões especiais | Criador do runbook | Última atualização | Contato para
 encaminhamento | |-------|-------|-------|-------|-------|-------|-------| | RUN001 |
 Para que serve este runbook? Qual é o resultado desejado? | Ferramentas | Permissões
 | Seu nome | 21-09-2022 | Nome para encaminhamento | ## Etapas 1. Primeira etapa 2.
 Segunda etapa
```

1. Se você não tiver um repositório de documentação ou uma wiki, crie um repositório de controle de versão em seu sistema de controle de versão.
2. Identifique um processo que não tenha um runbook. Um processo ideal é um que seja realizado quase regularmente, que tenha poucas etapas e que tenha falhas de baixo impacto.
3. No repositório de documentos, crie um rascunho de documento em Markdown usando o modelo. Preencha `Título do runbook` e os campos necessários em `Informações do runbook`.
4. Começando pela primeira etapa, preencha a seção `Etapas` do runbook.
5. Dê o runbook a um membro da equipe. Peça que o use para validar as etapas. Se algo estiver faltando ou não estiver claro, atualize o runbook.
6. Disponibilize o runbook em seu armazenamento interno de documentos. Depois, informe a sua equipe e outras partes interessadas.
7. Com o passar do tempo, você terá uma biblioteca de runbooks. À medida que essa biblioteca cresce, comece a trabalhar na automatização dos runbooks.

Nível de esforço do plano de implementação: baixo. O padrão mínimo para um runbook é um guia de texto detalhado. A automatização dos runbooks pode aumentar o esforço de implementação.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP02 Processos e procedimentos com proprietários identificados: os runbooks devem ter um proprietário responsável por mantê-los.
- OPS07-BP04 Usar manuais para investigar problemas: os runbooks e playbooks são semelhantes, com uma diferença importante: um runbook tem um resultado desejado. Em muitos casos, os runbooks são acionados depois que um playbook identifica uma causa raiz.
- OPS10-BP01 Usar um processo para gerenciamento de eventos, incidentes e problemas: os runbooks fazem parte de uma boa prática de gerenciamento de eventos, incidentes e problemas.
- OPS10-BP02 Ter um processo por alerta: os runbooks e playbooks devem ser usados para responder a alertas. Com o tempo, essas reações devem ser automatizadas.
- OPS11-BP04 Executar o gerenciamento de conhecimento: a manutenção dos runbooks é essencial para o gerenciamento de conhecimento.

Documentos relacionados:

- Como alcançar excelência operacional usando playbooks e runbooks automatizados
- AWS Systems Manager: trabalhar com runbooks
- Playbook para grandes migrações da AWS - Tarefa 4: Como melhorar runbooks de migração
- Como usar runbooks do AWS Systems Manager Automation para resolver tarefas operacionais

Vídeos relacionados:

- AWS re:Invent 2019: DIY guide to runbooks, incident reports, and incident response (SEC318-R1) (Guia DIY para runbooks, relatórios de incidentes e resposta a incidentes)
- How to automate IT Operations on AWS | Amazon Web Services (Como automatizar operações de TI na AWS | Amazon Web Services)
- Integrate Scripts into AWS Systems Manager (Integração de scripts no AWS Systems Manager)

Exemplos relacionados:

- AWS Systems Manager: demonstrações de automação
- AWS Systems Manager: runbook para restaurar um volume raiz usando o snapshot mais recente
- Criar um runbook de resposta a incidentes da AWS usando cadernos Jupyter e CloudTrail Lake
- Gitlab: runbooks
- Rubix: uma biblioteca de Python para criação de runbooks em cadernos Jupyter
- Como usar o gerador de documentos para criar um runbook personalizado
- Well-Architected Labs: automatização de operações com playbooks e runbooks

Serviços relacionados:

- AWS Systems Manager Automation

OPS07-BP04 Usar manuais para investigar problemas

Os manuais são guias detalhados usados para investigar incidentes. Quando incidentes ocorrem, os manuais são usados para investigar, definir o escopo do impacto e identificar a causa raiz. Os manuais são usados em diversos cenários, desde falhas em implantações até incidentes de segurança. Em muitos casos, os manuais identificam a causa raiz mitigada por um runbook. Os manuais são essenciais aos planos de resposta a incidentes de sua organização.

Um bom manual abrange vários aspectos principais. Ele guia o usuário, detalhadamente, ao longo do processo de descoberta. Considerando várias perspectivas, quais etapas devem ser seguidas para diagnosticar um incidente? Defina claramente no manual se são necessárias ferramentas especiais ou permissões elevadas. Ter um plano de comunicação para atualizar as partes interessadas sobre o status da investigação é essencial. Em situações em que a causa raiz ainda não foi identificada, o manual deve ter um plano de escalação. Se a causa raiz tiver sido identificada, o manual deverá indicar um runbook que descreva como resolvê-la. Os manuais devem ser armazenados em um local central e atualizados com frequência. Caso os manuais sejam usados para alertas específicos, forneça às equipes indicadores para o manual no alerta.

À medida que sua organização for amadurecendo, automatize seus manuais. Comece com manuais que abordem incidentes de baixo risco. Use scripts para automatizar as etapas de descoberta. Tenha runbooks complementares para mitigar as causas raízes comuns.

Resultado desejado: Sua organização tem manuais para incidentes comuns. Os manuais são armazenados em um local central e estão disponíveis para os membros da equipe. Os manuais são atualizados com frequência. São criados runbooks complementares para todas as causas raízes conhecidas.

Antipadrões comuns:

- Não há uma maneira padrão de investigar um incidente.
- Os membros da equipe precisam confiar na própria memória ou no conhecimento institucional para solucionar uma falha na implantação.
- Os novos membros da equipe aprendem a investigar os problemas por meio de tentativa e erro.
- As práticas recomendadas para a investigação dos problemas não são compartilhadas entre as equipes.

Benefícios de estabelecer esta prática recomendada:

- Os manuais impulsionam seus esforços para mitigar os incidentes.
- Diferentes membros da equipe podem usar o mesmo manual para identificar uma causa raiz de maneira consistente.
- As causas raízes conhecidas podem ter runbooks desenvolvidos para elas, o que acelera o tempo de recuperação.
- Os manuais permitem que os membros da equipe comecem a contribuir o quanto antes.
- As equipes podem escalar seus processos com manuais repetíveis.

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

Orientação para implementação

A maneira que você cria e usa os manuais depende da maturidade de sua organização. Se você é iniciante na nuvem, crie manuais no formato de texto em um repositório central de documentos. À medida que sua organização amadurecer, os manuais poderão passar a ser semiautomatizados com linguagens de script, como Python. Esses scripts podem ser executados em um caderno Jupyter para acelerar a descoberta. As organizações avançadas têm manuais totalmente automatizados para problemas comuns que são corrigidos automaticamente com runbooks.

Comece a criar seus manuais listando incidentes comuns que ocorrem com sua workload. Para começar, escolha manuais para incidentes com baixo risco e nos quais a causa raiz tenha sido restrita a poucos problemas. Quando você tiver manuais para os cenários mais simples, passe para cenários de alto risco ou cenários em que a causa raiz não seja bem conhecida.

Seus manuais em texto deverão ser automatizados à medida que sua organização amadurecer. Usando serviços, como o AWS Systems Manager Automations, um texto sem formatação pode ser transformado em automações. Essas automações podem ser executadas em sua workload para acelerar as investigações. Elas podem ser ativadas em resposta a eventos, o que reduz o tempo necessário para descobrir e resolver incidentes.

Os clientes podem usar o AWS Systems Manager Incident Manager para responder a incidentes. Esse serviço fornece uma interface única para fazer a triagem de incidentes, informar as partes interessas durante a descoberta e a mitigação e colaborar durante todo o incidente. Ele usa o AWS Systems Manager Automations para acelerar a detecção e a recuperação.

Exemplo de cliente

Um incidente na produção afetou a Loja UmaEmpresa. O engenheiro de plantão usou um manual para investigar o problema. À medida que foi avançando pelas etapas, ele manteve atualizadas as principais partes interessadas, identificadas no manual. O engenheiro identificou a causa raiz como uma condição de corrida em um serviço de back-end. Usando um runbook, o engenheiro reiniciou o serviço, colocando a Loja UmaEmpresa online novamente.

Etapas da implementação

Se você não tem um repositório de documentos, sugerimos criar um repositório de controle de versão para a biblioteca do manual. É possível criar os manuais usando o Markdown, que é compatível com a maioria dos sistemas de automação de manuais. Se você estiver iniciando do zero, use o modelo de exemplo de manual a seguir.

```
# Título do manual ## Informações do manual | ID do manual | Descrição |
 Ferramentas usadas | Permissões especiais | Autor do manual | Última atualização
 | Ponto de contato de escalação | Partes interessadas | Plano de comunicação |
 |-------|-------|-------|-------|-------|-------|-------|-------|-------| | RUN001 |
 Para que é este manual? Ele é usado para qual incidente? | Ferramentas | Permissões
 | Seu nome | 21/9/2022 | Nome para escalação | Nome da parte interessada | Como as
 atualizações serão comunicadas durante a investigação? | ## Etapas 1. Etapa um 2.
 Etapa dois
```

1. Se você não tiver um repositório de documentos ou uma wiki, crie um repositório de controle de versão para seus manuais no sistema de controle de versão.
2. Identifique um problema comum que requer investigação. Ele deve ser um cenário em que a causa raiz esteja limitada a poucos problemas e a resolução seja de baixo risco.
3. Usando o modelo do Markdown, preencha a seção `Nome do manual` e os campos em `Informações do manual`.
4. Preencha as etapas de resolução de problemas. Seja o mais claro possível sobre quais ações devem ser executadas ou quais áreas devem ser investigadas.
5. Dê o manual a um membro da equipe e peça para essa pessoa analisá-lo a fim de validá-lo. Caso algo esteja faltando ou não esteja claro, atualize o manual.
6. Publique o manual no repositório de documentos e informe sua equipe e as partes interessadas.
7. Essa biblioteca de manuais crescerá à medida que você adicionar outros manuais. Quando você tiver vários manuais, comece a automatizá-los usando ferramentas como o AWS Systems Manager Automations a fim de manter a automação e os manuais sincronizados.

Nível de esforço do plano de implementação: Baixo. Os manuais devem ser documentos de texto armazenados em um local central. Organizações mais consolidadas passarão a automatizar os respectivos manuais.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP02 Processos e procedimentos com proprietários identificados: os manuais devem ter um proprietário responsável por mantê-los.
- OPS07-BP03 Usar runbooks para realizar procedimentos: os runbooks e os manuais são semelhantes, com uma diferença importante: um runbook tem um resultado desejado. Em muitos casos, os runbooks são usados quando um manual identifica uma causa raiz.

- OPS10-BP01 Usar um processo para gerenciamento de eventos, incidentes e problemas: os manuais fazem parte de uma boa prática de gerenciamento de eventos, incidentes e problemas.
- OPS10-BP02 Ter um processo por alerta: os runbooks e manuais devem ser usados para responder a alertas. Com o tempo, essas reações devem ser automatizadas.
- OPS11-BP04 Executar o gerenciamento de conhecimento: a manutenção dos manuais é essencial para o gerenciamento de conhecimento.

Documentos relacionados:

- Achieving Operational Excellence using automated playbook and runbook (Como alcançar excelência operacional usando manuais e runbooks automatizados)
- AWS Systems Manager: Working with runbooks (AWS Systems Manager: trabalho com runbooks)
- Use AWS Systems Manager Automation runbooks to resolve operational tasks (Usar runbooks do AWS Systems Manager Automation para resolver tarefas operacionais)

Vídeos relacionados:

- AWS re:Invent 2019: DIY guide to runbooks, incident reports, and incident response (SEC318-R1) (Guia DIY para runbooks, relatórios de incidentes e resposta a incidentes)
- AWS Systems Manager Incident Manager - AWS Virtual Workshops (AWS Systems Manager Incident Manager - workshops virtuais da AWS)
- Integrate Scripts into AWS Systems Manager (Integração de scripts no AWS Systems Manager)

Exemplos relacionados:

- AWS Customer Playbook Framework (Framework do manual do cliente daAWS)
- AWS Systems Manager: Automation walkthroughs (AWS Systems Manager: demonstrações de automação)
- Building an AWS incident response runbook using Jupyter notebooks and CloudTrail Lake (Criar um runbook de resposta a incidentes da AWS usando cadernos Jupyter e o CloudTrail Lake)
- Rubix – A Python library for building runbooks in Jupyter Notebooks (Rubix: uma biblioteca de Python para criação de runbooks em cadernos Jupyter)
- Using Document Builder to create a custom runbook (Como usar o gerador de documentos para criar um runbook personalizado)

- Well-Architected Labs: Automating operations with Playbooks and Runbooks (Well-Architected Labs: automatização de operações com manuais e runbooks)
- Well-Architected Labs: Incident response playbook with Jupyter (Well-Architected Labs: manual de resposta a incidentes com o Jupyter)

Serviços relacionados:

- AWS Systems Manager Automation
- AWS Systems Manager Incident Manager

OPS07-BP05 Tomar decisões embasadas para implantar sistemas e alterações

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.

Uma estratégia pre-mortem é um exercício em que uma equipe simula uma falha para desenvolver estratégias de mitigação. Use estratégias pre-mortem para prever falhas e criar procedimentos, quando apropriado. 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.

Antipadrões comuns:

- Decidir implantar uma carga de trabalho sem entender os riscos de segurança presentes na carga de trabalho.
- Decidir implantar uma carga de trabalho sem entender se ela está em conformidade com sua governança e seus padrões.
- Decidir implantar uma carga de trabalho sem entender se sua equipe pode oferecer suporte a ela.
- Decidir implantar uma carga de trabalho sem entender como ela beneficia a organização.

Benefícios do estabelecimento desta prática recomendada: Ter membros da equipe qualificados possibilita o suporte eficaz da sua carga de trabalho.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Tomar decisões embasadas para implantar workloads e alterações: avalie os recursos da equipe para apoiar a workload e a conformidade da workload 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 e tome decisões informadas.

Operar

Perguntas

- OPS 8 Como você compreende a integridade da sua carga de trabalho?
- OPS 9 Como você compreende a integridade de suas operações?
- OPS 10 Como você gerencia os eventos de carga de trabalho e operações?

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.

Práticas recomendadas

- OPS08-BP01 Identificar os indicadores-chave de performance
- OPS08-BP02 Definir as métricas da workload
- OPS08-BP03 Coletar e analisar as métricas da workload
- OPS08-BP04 Estabelecer as linhas de base das métricas da workload
- OPS08-BP05 Aprender os padrões esperados das atividades da workload
- OPS08-BP06 Alertar quando os resultados da workload estiverem em risco
- OPS08-BP07 Alertar quando forem detectadas anomalias na workload
- OPS08-BP08 Validar a obtenção de resultados e a eficácia dos KPIs e das métricas

OPS08-BP01 Identificar 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.

Antipadrões comuns:

- A liderança de negócios pergunta a você sobre o sucesso de uma carga de trabalho atendendo às necessidades empresariais, mas não tem um quadro de referência para determinar o sucesso.
- Você não consegue determinar se a aplicação comercial pronta para uso que você opera para a organização é econômica.

Benefícios do estabelecimento desta prática recomendada: Ao identificar os indicadores-chave de performance, você permite alcançar resultados empresariais como teste da integridade e do sucesso da sua carga de trabalho.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Identificar os indicadores-chave de performance: identifique os indicadores-chave de performance (KPIs) com base nos resultados desejados dos negócios e dos clientes. Avalie os KPIs para determinar o sucesso da carga de trabalho.

OPS08-BP02 Definir as métricas da workload

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.

Você deve enviar os dados de log para um serviço como o CloudWatch Logs e gerar métricas a partir das observações do conteúdo do log necessário.

O CloudWatch tem recursos especializados, como Amazon CloudWatch Insights para .NET e SQL Server e Container Insights , que podem ajudar você ao identificar e configurar as principais métricas, logs e alarmes em seus recursos de aplicativos e pilha de tecnologia especificamente com suporte.

Antipadrões comuns:

- Você definiu métricas padrão, não associadas a nenhum KPI nem adaptadas a nenhuma workload.
- Os cálculos de métricas apresentam erros que produzirão resultados inválidos.
- Não há nenhuma métrica definida para sua carga de trabalho.
- Você só mede a disponibilidade.

Benefícios do estabelecimento desta prática recomendada: Ao definir e avaliar métricas de carga de trabalho, você pode determinar a integridade da sua carga de trabalho e medir a obtenção dos resultados de negócios.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Definir as métricas da workload: defina as métricas da workload para medir o alcance dos KPIs. Defina métricas de carga de trabalho para medir a sua integridade e a de seus componentes individuais. Avalie as métricas para determinar se a carga de trabalho está alcançando os resultados desejados e para entender a sua integridade.
  - Publique métricas personalizadas.
  - Pesquisa e filtragem de dados de log
  - Referência de métricas e de dimensões do Amazon CloudWatch

Recursos

Documentos relacionados:

- Referência de métricas e de dimensões do Amazon CloudWatch
- Publicar métricas personalizadas
- Pesquisa e filtragem de dados de log

OPS08-BP03 Coletar e analisar as métricas da workload

Faça revisões proativas regulares das métricas para identificar tendências e determine onde as respostas apropriadas são necessárias.

Agregue os dados de log da aplicação, dos componentes da workload, dos serviços e das chamadas de API para um serviço como o CloudWatch Logs. Gere métricas a partir de observações do conteúdo de log necessário para permitir insights sobre a performance de atividades de operações.

Na AWS, é possível analisar as métricas da workload e identificar problemas operacionais usando os recursos de machine learning do Amazon DevOps GuruO AWS DevOps Guru fornece notificação de problemas operacionais com recomendações direcionadas e proativas para resolver problemas e manter a integridade da aplicação.

No modelo de responsabilidade compartilhada da AWS, partes do monitoramento são entregues por meio do AWS Health Dashboard. O painel fornece alertas e orientação de remediação quando a AWS apresenta eventos que podem afetar você. Os clientes com assinaturas do Business e Enterprise Support também obtêm acesso à API do AWS Health, permitindo a integração com seus sistemas de gerenciamento de eventos.

Na AWS, você pode exportar seus dados de log para o Amazon S3 ou enviar logs diretamente to Amazon S3 para armazenamento de longo prazo. Com o uso do AWS Glue, você pode descobrir e preparar seus dados de log no Amazon S3 para análises, armazenando metadados associados no AWSAWS Glue Data Catalog. Amazon Athena, por meio da integração nativa com o AWS Glue, pode ser usado para analisar dados de log, consultando-os com o SQL padrão. Usando uma ferramenta de business intelligence, como o Amazon QuickSight você pode visualizar, explorar e analisar seus dados.

Uma solução alternativa seria usar o Amazon OpenSearch Service e os painéis do OpenSearch para coletar, analisar e exibir logs na AWS em várias contas e Regiões da AWS.

Antipadrões comuns:

- A equipe de design de rede solicita as taxas de utilização de largura de banda de rede atuais. Você fornece as métricas atuais, a utilização da rede é de 35%. Elas reduzem a capacidade do circuito como uma medida de economia de custos, causando problemas de conectividade generalizados, pois sua medição de ponto no tempo não reflete a tendência nas taxas de utilização.
- O roteador falhou. Ele está registrando erros de memória não críticos com frequência cada vez maior, até sua falha completa. Você não detectou essa tendência e, como resultado, não substituiu a memória com falha antes que o roteador causasse uma interrupção no serviço.

Benefícios do estabelecimento desta prática recomendada: Ao coletar e analisar as métricas de carga de trabalho, você compreende a integridade da sua carga de trabalho e pode obter

informações sobre tendências que podem afetar sua carga de trabalho ou a obtenção de seus resultados de negócios.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Coletar e analisar métricas da workload: execute revisões proativas regulares de métricas para identificar tendências e determinar quando as respostas apropriadas são necessárias.
  - Uso de métricas do Amazon CloudWatch
  - Referência de métricas e de dimensões do Amazon CloudWatch
  - Collect metrics and logs from Amazon EC2 instances and on-premises servers with the CloudWatch Agent (Coletar métricas e logs das instâncias do Amazon EC2 e de servidores on-premises com o agente do CloudWatch)

Recursos

Documentos relacionados:

- Amazon Athena
- Referência de métricas e de dimensões do Amazon CloudWatch
- Amazon DevOps Guru
- AWS Glue
- AWSAWS Glue Data Catalog
- Amazon OpenSearch Service
- AWS Health Dashboard
- Amazon QuickSight
- Collect metrics and logs from Amazon EC2 instances and on-premises servers with the CloudWatch Agent (Coletar métricas e logs das instâncias do Amazon EC2 e de servidores on-premises com o agente do CloudWatch)
- Uso de métricas do Amazon CloudWatch

OPS08-BP04 Estabelecer as linhas de base das métricas da workload

Estabeleça as linhas de base das métricas para fornecer os valores esperados como base para a comparação e a identificação de componentes com performance inferior e superior. Identificar limites para melhoria, investigação e intervenção.

Antipadrões comuns:

- Um servidor está sendo executado com 95% de utilização da CPU. Será perguntado se isso é bom ou ruim. A utilização da CPU nesse servidor não foi usada como base, portanto, você não tem ideia se isso é bom ou ruim.

Benefícios do estabelecimento desta prática recomendada: Ao definir valores de métrica de linha de base, você pode avaliar valores de métrica atuais e tendências de métrica para determinar se a ação é necessária.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Estabelecer as linhas de base para as métricas da workload: estabeleça as linhas de base das métricas da workload para fornecer os valores esperados como uma base de comparação.
  - Criação de alarmes do Amazon CloudWatch

Recursos

Documentos relacionados:

- Criação de alarmes do Amazon CloudWatch

OPS08-BP05 Aprender os padrões esperados das atividades da workload

Estabeleça padrões de atividade de carga de trabalho para identificar comportamentos anômalos para que você possa responder adequadamente, se necessário.

O CloudWatch por meio da Detecção de anomalias do CloudWatch aplica algoritmos estatísticos e de machine learning para gerar uma variedade de valores esperados que representam o comportamento normal da métrica.

Amazon DevOps Guru pode ser usado para identificar comportamento anômalo por meio da correlação de eventos, da análise do log e da aplicação de machine learning para analisar a telemetria da workload. Quando são detectados comportamentos inesperados, ele fornece as métricas e os eventos relacionados com recomendações para resolver o comportamento.

Antipadrões comuns:

- Você está revisando os logs de utilização da rede e verá que a utilização da rede aumentou entre 11h30 e 13h30 e novamente das 16h30 às 18h. Você não sabe se isso deve ser considerado normal ou não.
- Seus servidores web reinicializam todas as noites às 3h. Você não sabe se esse é um comportamento esperado.

Benefícios do estabelecimento desta prática recomendada: Ao aprender padrões de comportamento, você pode reconhecer comportamentos inesperados e tomar medidas, se necessário.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Estabelecer os padrões esperados das atividades da workload: estabeleça os padrões das atividades da workload para determinar quando o comportamento está fora dos valores esperados, para que seja possível responder adequadamente, se necessário.

Recursos

Documentos relacionados:

- Amazon DevOps Guru
- Detecção de anomalias do CloudWatch

OPS08-BP06 Alertar quando os resultados da workload estiverem em risco

Emita um alerta quando os resultados da carga de trabalho estiverem em risco, para que você possa responder adequadamente, se necessário.

em condições ideais, você identificou anteriormente um limite de métrica sobre o qual é capaz de emitir alarmes ou um evento que você pode usar para acionar uma resposta automatizada.

No AWS, você pode usar o Amazon CloudWatch Synthetics para criar scripts canário para monitorar os seus endpoints e APIs executando as mesmas ações que seus clientes. A telemetria gerada e o insight obtido podem permitir que você identifique problemas antes que causem impacto nos clientes.

Você também pode usar o CloudWatch Logs Insights para pesquisar e analisar interativamente seus dados de log usando uma linguagem de consulta específica. O CloudWatch Logs Insights descobre campos em logs automaticamente dos serviços da AWS e dos eventos de log personalizados em JSON. Ele faz o dimensionamento de acordo com o volume de logs e a complexidade das consultas e oferece respostas em segundos, ajudando você a procurar os fatores que contribuem para um incidente.

Antipadrões comuns:

- Você não tem conectividade de rede. Ninguém está ciente. Ninguém está tentando identificar o motivo ou tomando medidas para restaurar a conectividade.
- Após a aplicação de um patch, as instâncias persistentes se tornaram indisponíveis, prejudicando os usuários. Seus usuários abriram casos de suporte. Ninguém foi notificado. Ninguém está realizando ações.

Benefícios do estabelecimento desta prática recomendada: Ao identificar que os resultados de negócios estão em risco e alertar sobre ações a serem tomadas, você tem a oportunidade de evitar ou reduzir o impacto de um incidente.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Alertar quando os resultados da workload estão em risco: emita um alerta quando os resultados da workload estiverem em risco para que você possa responder adequadamente, se necessário.
  - O que é o Amazon CloudWatch Events?
  - Criação de alarmes do Amazon CloudWatch
  - Invocar funções do Lambda usando notificações do Amazon SNS

Recursos

Documentos relacionados:

- Amazon CloudWatch Synthetics

- CloudWatch Logs Insights
- Criação de alarmes do Amazon CloudWatch
- Invocar funções do Lambda usando notificações do Amazon SNS
- O que é o Amazon CloudWatch Events?

OPS08-BP07 Alertar quando forem detectadas anomalias na workload

Emita um alerta quando forem detectadas anomalias na carga de trabalho, para que você possa responder adequadamente, se necessário.

sua análise das métricas da carga de trabalho ao longo do tempo pode estabelecer padrões de comportamento que você pode quantificar suficientemente para definir um evento ou gerar um alarme em resposta.

Uma vez treinado, o recurso Detecção de anomalias do CloudWatch pode ser usado para gerar alarmes sobre anomalias detectadas ou pode fornecer valores esperados sobrepostos em um gráfico de dados de métricas para comparação contínua.

Antipadrões comuns:

- As vendas do site de varejo aumentaram drasticamente de forma repentina; Ninguém está ciente. Ninguém está tentando identificar o que levou a esse pico. Ninguém está realizando ações para garantir experiências de qualidade para o cliente sob a carga adicional.
- Após a aplicação de um patch, seus servidores persistentes estão reinicializando com frequência, prejudicando os usuários. Normalmente, os servidores reinicializam até três vezes, mas não mais. Ninguém está ciente. Ninguém está tentando identificar por que isso está acontecendo.

Benefícios do estabelecimento desta prática recomendada: Com a compreensão dos padrões de comportamento da workload, é possível identificar comportamentos inesperados e tomar medidas, se necessário.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Alertar quando são detectadas anomalias da workload: emita um alerta quando anomalias da workload forem detectadas para que seja possível responder adequadamente, se necessário.
  - O que é o Amazon CloudWatch Events?

- Criação de alarmes do Amazon CloudWatch
- Invocar funções do Lambda usando notificações do Amazon SNS

Recursos

Documentos relacionados:

- Criação de alarmes do Amazon CloudWatch
- Detecção de anomalias do CloudWatch
- Invocar funções do Lambda usando notificações do Amazon SNS
- O que é o Amazon CloudWatch Events?

OPS08-BP08 Validar a obtenção de resultados e a eficácia dos KPIs e das 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.

A AWS também é compatível com sistemas de análise de log de terceiros e com ferramentas de inteligência de negócios por meio das APIs e SDKs de serviços da AWS (por exemplo, Grafana, Kibana e Logstash).

Antipadrões comuns:

- O tempo de resposta da página nunca foi considerado um colaborador para a satisfação do cliente. Você nunca estabeleceu uma métrica ou um limite para o tempo de resposta da página. Seus clientes estão reclamando sobre lentidão.
- Você não está atingindo seus objetivos mínimos de tempo de resposta. Como um esforço para melhorar o tempo de resposta, você aumentou a escala vertical dos servidores de aplicações. Agora você está excedendo as metas de tempo de resposta por uma margem significativa e também tem uma capacidade significativa não utilizada pela qual está pagando.

Benefícios do estabelecimento desta prática recomendada: Ao analisar e revisar os KPIs e as métricas, você entende como sua workload oferece suporte à obtenção dos resultados dos negócios e pode identificar onde é necessário melhorar para atingir suas metas de negócios.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Validar a obtenção dos resultados e a eficácia dos KPIs e das métricas: crie uma visão de nível empresarial das operações da workload para ajudá-lo a determinar se você está atendendo às necessidades e a identificar áreas que precisam ser aprimoradas para atingir metas empresariais. Valide a eficácia dos KPIs e métricas e revise-os, se necessário.
  - Uso de painéis do Amazon CloudWatch
  - O que é análise de log?

Recursos

Documentos relacionados:

- Uso de painéis do Amazon CloudWatch
- O que é análise de log?

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.

Práticas recomendadas

- OPS09-BP01 Identificar os indicadores-chave de performance
- OPS09-BP02 Definir as métricas das operações
- OPS09-BP03 Coletar e analisar as métricas de operações
- OPS09-BP04 Estabelecer linhas de base das métricas de operações
- OPS09-BP05 Aprender os padrões esperados de atividades das operações
- OPS09-BP06 Alertar quando os resultados das operações estão em risco
- OPS09-BP07 Alertar quando são detectadas anomalias nas operações
- OPS09-BP08 Validar a obtenção de resultados e a eficácia dos KPIs e das métricas

OPS09-BP01 Identificar os indicadores-chave de performance

Identifique os indicadores-chave de performance (KPIs) com base nos resultados dos 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.

Antipadrões comuns:

- A liderança de negócios pergunta se as operações são bem-sucedidas na realização de metas empresariais, mas não tem um quadro de referência para determinar o sucesso.
- Não é possível determinar se as janelas de manutenção têm impacto nos resultados de negócios.

Benefícios do estabelecimento desta prática recomendada: Ao identificar os indicadores-chave de performance, você permite alcançar resultados de negócios, assim como o teste da integridade e do sucesso das suas operações.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Identificar os indicadores-chave de performance: identifique os indicadores-chave de performance (KPIs) com base nos resultados desejados dos negócios e dos clientes. Avalie KPIs para determinar o sucesso das operações.

OPS09-BP02 Definir as métricas das 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.

Antipadrões comuns:

- As métricas de operações são baseadas no que a equipe considera razoável.
- Os cálculos de métricas apresentam erros que produzirão resultados incorretos.
- Não há nenhuma métrica definida para suas atividades operacionais.

Benefícios do estabelecimento desta prática recomendada: Ao definir e avaliar métricas de operações, você pode determinar a integridade de suas atividades de operações e medir a obtenção de resultados de negócios.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Definir as métricas das operações: defina as métricas das operações para medir a realização dos KPIs. Defina as métricas de operações para medir a integridade das operações e de suas atividades. Avalie as métricas para determinar se as operações estão alcançando os resultados desejados e para entender a integridade das operações.
  - Publique métricas personalizadas.
  - Pesquisa e filtragem de dados de log
  - Referência de métricas e de dimensões do Amazon CloudWatch

Recursos

Documentos relacionados:

- AWS Answers: Centralized Logging (Resposta da AWS: registro em log centralizado)
- Referência de métricas e de dimensões do Amazon CloudWatch
- Identificar e responder a alterações no estado do pipeline com o Amazon CloudWatch Events
- Publicar métricas personalizadas
- Pesquisa e filtragem de dados de log

Vídeos relacionados:

- Build a monitoring plan

OPS09-BP03 Coletar e analisar as métricas de operações

Faça revisões proativas regulares das métricas para identificar tendências e determine onde as respostas apropriadas são necessárias.

Agregue os dados de log da execução de suas atividades de operações e chamadas de API de operações em um serviço como o CloudWatch Logs. Gere métricas a partir de observações do conteúdo de log necessário para obter insights sobre a performance das atividades de operações.

Na AWS, você pode exportar seus dados de log para o Amazon S3 ou enviar logs diretamente to Amazon S3 para armazenamento de longo prazo. Com o uso do AWS Glue, você pode descobrir e preparar seus dados de log no Amazon S3 para análises, armazenando metadados associados no AWSAWS Glue Data Catalog. Amazon Athena, por meio da integração nativa com o AWS Glue, pode ser usado para analisar dados de log, consultando-os com o SQL padrão. Usando uma ferramenta de business intelligence, como o Amazon QuickSight você pode visualizar, explorar e analisar seus dados.

Antipadrões comuns:

- A entrega consistente de novos recursos é considerada um indicador-chave de performance. Não há um método para medir a frequência com que as implantações ocorrem.
- Você registra implantações, implantações revertidas, patches e patches revertidos para rastrear suas atividades operacionais, mas ninguém analisa as métricas.
- Você tem um objetivo de tempo de recuperação para restaurar um banco de dados perdido em 15 minutos, que foi definido quando o sistema foi implantado e não tinha usuários. Agora, você tem milhares de usuários e está em operação há dois anos. Uma restauração recente levou mais de duas horas. Isso não foi registrado e ninguém está ciente.

Benefícios do estabelecimento desta prática recomendada: Ao coletar e analisar as métricas de operações, você entende a integridade das operações e pode obter insights sobre as tendências que podem afetar as operações ou a obtenção dos resultados de negócios.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Coletar e analisar as métricas de operações: execute análises regulares e proativas das métricas para identificar tendências e determinar quando respostas apropriadas são necessárias.
  - Uso de métricas do Amazon CloudWatch
  - Referência de métricas e de dimensões do Amazon CloudWatch
  - Coletar métricas e logs das instâncias do Amazon EC2 e de servidores on-premises com o agente do CloudWatch

Recursos

Documentos relacionados:

- Amazon Athena
- Referência de métricas e de dimensões do Amazon CloudWatch
- Amazon QuickSight
- AWS Glue
- AWSAWS Glue Data Catalog
- Coletar métricas e logs das instâncias do Amazon EC2 e de servidores on-premises com o agente do CloudWatch
- Uso de métricas do Amazon CloudWatch

OPS09-BP04 Estabelecer 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.

Antipadrões comuns:

- Foi perguntado a você qual é o tempo esperado para implantar. Você não mediu o tempo necessário para a implantação e não consegue determinar o tempo esperado.
- Foi perguntado a você quanto tempo leva para se recuperar de um problema com os servidores de aplicativos. Você não tem informações sobre o tempo de recuperação a partir do primeiro contato com o cliente. Você não tem informações sobre o tempo de recuperação a partir da primeira identificação de um problema por meio do monitoramento.
- Foi perguntado a você quantos funcionários de suporte são necessários durante o fim de semana. Você não tem ideia de quantos casos de suporte são realizados normalmente durante um fim de semana e não pode fornecer uma estimativa.
- Você tem um objetivo de tempo de recuperação para restaurar bancos de dados perdidos em 15 minutos, que foi definido quando o sistema foi implantado e não tinha usuários. Agora, você tem milhares de usuários e está em operação há dois anos. Você não tem informações sobre como o tempo de restauração foi alterado para seu banco de dados.

Benefícios do estabelecimento desta prática recomendada: Ao definir valores de métrica de linha de base, você pode avaliar valores de métrica atuais e tendências de métrica para determinar se a ação é necessária.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Aprender os padrões esperados das atividades da workload: estabeleça os padrões das atividades da workload para determinar quando o comportamento está fora dos valores esperados, para que seja possível responder adequadamente, se necessário.

OPS09-BP05 Aprender os padrões esperados de atividades das operações

Estabeleça padrões de atividades de operações para identificar atividades anômalas para poder responder adequadamente, se necessário.

Antipadrões comuns:

- A taxa de falhas de implantação aumentou substancialmente recentemente. Você aborda cada uma das falhas de forma independente. Você não percebe que as falhas correspondem a implantações de um novo funcionário que não está familiarizado com o sistema de gerenciamento de implantação.

Benefícios do estabelecimento desta prática recomendada: Ao aprender os padrões de comportamento, você pode reconhecer comportamentos inesperados e tomar medidas, se necessário.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Aprender os padrões esperados das atividades da workload: estabeleça os padrões das atividades da workload para determinar quando o comportamento está fora dos valores esperados, para que seja possível responder adequadamente, se necessário.

OPS09-BP06 Alertar quando os resultados das operações estão em risco

Sempre que os resultados da operação estiverem em risco, um alerta deve ser gerado e acionado. Os resultados das operações são qualquer atividade compatível com uma workload em produção.

Isso inclui tudo, desde a implantação de novas versões de aplicações até a recuperação de uma interrupção. Os resultados das operações devem ser tratados com a mesma importância dos resultados empresariais.

As equipes de software devem identificar as principais métricas e atividades da operação e criar alertas para elas. Os alertas devem ser enviados em tempo hábil e levar a ações concretas. Se um alerta for criado, deverá ser incluída uma referência para um runbook ou manual correspondente. Os alertas sem uma ação correspondente podem levar a um excesso de alertas.

Resultado desejado: quando as atividades das operações estão em risco, são enviados alertas para promover uma ação. Os alertas contêm contexto sobre por que o alerta está sendo criado e indicam um manual para investigação ou um runbook para mitigação. Quando possível, os runbooks são automatizados e as notificações são enviadas.

Antipadrões comuns:

- Você está investigando um incidente e os casos de suporte estão sendo arquivados. Os casos de suporte estão infringindo o Acordo de Serviço (SLA), mas nenhum alerta está sendo criado.
- Uma implantação na produção agendada para a meia-noite está atrasada devido a modificações de última hora no código. Nenhum alerta foi criado e a implantação é adiada.
- Uma interrupção da produção ocorre, mas não é enviado nenhum alerta.
- O tempo da implantação constantemente não cumpre o tempo estimado. Nenhuma ação é realizada para investigar.

Benefícios de estabelecer esta prática recomendada:

- Alertar quando os resultados das operações estiverem em risco aumenta sua capacidade de comportar sua workload, ao se antecipar aos problemas.
- Os resultados empresariais são melhorados devido a resultados operacionais íntegros.
- A detecção e correção dos problemas das operações são melhorados.
- A integridade operacional geral é melhorada.

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

Orientação para implementação

Os resultados das operações devem ser definidos antes de você poder alertar sobre eles. Comece definindo quais atividades das operações são mais importantes para sua organização. É implantar

na produção em menos de duas horas ou responder a um caso de suporte em determinado tempo? Sua organização deve definir as principais atividades de operações e como elas devem ser medidas, para que possam ser monitoradas, aprimoradas e alertadas. Você precisa de um local central em que a telemetria de operações e workload seja armazenada e analisada. O mesmo mecanismo deverá poder criar um alerta quando o resultado de uma operação estiver em risco.

Exemplo de cliente

Um alarme do CloudWatch foi acionado durante uma implantação de rotina na Loja UmaEmpresa. O tempo útil para a implantação foi violado. O Amazon EventBridge criou um OpsItem no AWS Systems Manager OpsCenter. A equipe de operações da nuvem usou um manual para investigar o problema e identificou que uma mudança no esquema estava levando mais tempo do que o esperado. Ela alertou o desenvolvedor de plantão e continuou a monitorar a implantação. Depois que a implantação foi concluída, a equipe de operações da nuvem resolveu o OpsItem. A equipe fará uma análise após a conclusão do incidente.

Etapas da implementação

1. Se você não identificou os KPIs, as métricas e as atividades da operação, trabalhe na implementação das práticas recomendadas anteriores a essa questão (de OPS09-BP01 a OPS09-BP05).
   - Clientes do AWS Support com Enterprise Support podem solicitar o workshop de KPI de operações com seu gerente de conta técnico. Esse workshop colaborativo ajuda a definir os KPIs e as métricas das operações de forma alinhada às metas empresariais, fornecidos sem custo adicional. Entre em contato com seu gerente de conta técnico para saber mais.
2. Depois de estabelecer as atividades, os KPIs e as métricas das operações, configure alertas em sua plataforma de observabilidade. Os alertas devem ter uma ação associada a eles, como um manual ou um runbook. Os alertas sem uma ação devem ser evitados.
3. Ao longo do tempo, você deve avaliar as métricas, KPIs e atividades das operações a fim de identificar áreas para melhoria. Colete feedback em runbooks e manuais dos operadores visando identificar áreas para melhoria ao responder a alertas.
4. Os alertas devem incluir um mecanismo para sinalizá-los como falso positivo. Isso deve levar a uma análise dos limites das métricas.

Nível de esforço do plano de implementação: médio. Há várias práticas recomendadas que devem ser aplicadas antes de implementar essa prática recomendada. Depois de identificar as atividades e definir os KPIs das operações, estabeleça alertas.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP03 Atividades de operações com proprietários identificados responsáveis pela performance: todas as atividades e os resultados da operação devem ter um proprietário identificado como responsável. Essa é a pessoa que deverá ser alertada quando os resultados estiverem em risco.
- OPS03-BP02 Os membros da equipe estão capacitados para executar ações quando os resultados estão em risco.: quando os alertas forem criados, sua equipe deverá ter autoridade para atuar a fim de corrigir o problema.
- OPS09-BP01 Identificar os indicadores-chave de performance: os alertas com relação aos resultados das operações começam com a identificação dos KPIs das operações.
- OPS09-BP02 Definir as métricas das operações: estabeleça essa prática recomendada antes de começar a gerar alertas.
- OPS09-BP03 Coletar e analisar as métricas de operações: é necessário coletar centralmente as métricas das operações para criar alertas.
- OPS09-BP04 Estabelecer linhas de base das métricas de operações: as referências de base das métricas de operações permitem ajustar os alertas e evitar o excesso de alertas.
- OPS09-BP05 Aprender os padrões esperados de atividades das operações: é possível melhorar a precisão de seus alertas compreendendo os padrões de atividades dos eventos de operações.
- OPS09-BP08 Validar a obtenção de resultados e a eficácia dos KPIs e das métricas: avalie o cumprimento dos resultados das operações para garantir a validade dos KPIs e das métricas.
- OPS10-BP02 Ter um processo por alerta: todos os alertas devem ter um runbook ou manual associado e fornecer contexto para a pessoa que recebe o alerta.
- OPS11-BP02 Executar análise pós-incidente: realize uma análise pós-incidente depois do alerta para identificar áreas para melhoria.

Documentos relacionados:

- AWS Deployment Pipelines Reference Architecture: Application Pipeline Architecture (Arquitetura de referência de pipelines de implantação da AWS: arquitetura de pipeline de aplicação)
- GitLab: Getting Started with Agile / DevOps Metrics (GitLab conceitos básicos do Agile/métricas de DevOps)

Vídeos relacionados:

- Aggregate and Resolve Operational Issues Using AWS Systems Manager OpsCenter (Agregue e resolva problemas operacionais usando o AWS Systems Manager OpsCenter)
- Integrate AWS Systems Manager OpsCenter with Amazon CloudWatch Alarms (Integre o AWS Systems Manager OpsCenter com alarmes do Amazon CloudWatch)
- Integrate Your Data Sources into AWS Systems Manager OpsCenter Using Amazon EventBridge (Integre suas fontes de dados ao AWS Systems Manager OpsCenter usando o Amazon EventBridge)

Exemplos relacionados:

- Automate remediation actions for Amazon EC2 notifications and beyond using Amazon EC2 Systems Manager Automation and AWS Health (Automatize ações de correção para notificações do Amazon EC2 e além usando o Amazon EC2 Systems Manager Automation e o AWS Health)
- AWS Management and Governance Tools Workshop - Operations 2022 (Workshop de ferramentas de gerenciamento e governança da AWS: Operações de 2022)
- Ingesting, analyzing, and visualizing metrics with DevOps Monitoring Dashboard on AWS (Ingerir, analisar e visualizar métricas com o painel de monitoramento de DevOps na AWS)

Serviços relacionados:

- Amazon EventBridge
- AWS Support Proactive Services - Operations KPI Workshop (Serviços proativos do AWS Support: workshop de KPI de operações)
- AWS Systems Manager OpsCenter
- Eventos do CloudWatch

OPS09-BP07 Alertar quando são detectadas anomalias nas operações

Emita um alerta quando forem detectadas anomalias de operações para que você possa responder adequadamente, se necessário.

Sua análise das métricas de operações ao longo do tempo pode estabelecer padrões de comportamento que você pode quantificar suficientemente para definir um evento ou gerar um alarme em resposta.

Uma vez treinado, o recurso Detecção de anomalias do CloudWatch pode ser usado para gerar alarmes sobre anomalias detectadas ou pode fornecer valores esperados sobrepostos em um gráfico de dados de métricas para comparação contínua.

Amazon DevOps Guru pode ser usado para identificar comportamento anômalo por meio da correlação de eventos, da análise do log e da aplicação de machine learning para analisar a telemetria da workload. O insights obtidos são apresentados com os dados e as recomendações relevantes.

Antipadrões comuns:

- Você está aplicando um patch à sua frota de instâncias. Você testou o patch com êxito no ambiente de teste. O patch está falhando para uma grande porcentagem de instâncias em sua frota. Você não faz nada.
- Você percebe que há implantações a partir da sexta-feira no fim do dia. Sua organização tem janelas de manutenção predefinidas às terças e quintas-feiras. Você não faz nada.

Benefícios do estabelecimento desta prática recomendada: Ao compreender os padrões de comportamento das operações, é possível identificar comportamentos inesperados e tomar medidas, se necessário.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Alertar quando são detectadas anomalias nas operações: emita um alerta quando forem detectadas anomalias nas operações para poder responder adequadamente, se necessário.
  - O que é o Amazon CloudWatch Events?
  - Criação de alarmes do Amazon CloudWatch
  - Invocar funções do Lambda usando notificações do Amazon SNS

Recursos

Documentos relacionados:

- Amazon DevOps Guru
- Detecção de anomalias do CloudWatch
- Criação de alarmes do Amazon CloudWatch

- Identificar e responder a alterações no estado do pipeline com o Amazon CloudWatch Events
- Invocar funções do Lambda usando notificações do Amazon SNS
- O que é o Amazon CloudWatch Events?

OPS09-BP08 Validar a obtenção de resultados e a eficácia dos KPIs e das 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.

A AWS também é compatível com sistemas de análise de log de terceiros e com ferramentas de inteligência de negócios por meio das APIs e SDKs de serviços da AWS (por exemplo, Grafana, Kibana e Logstash).

Antipadrões comuns:

- A frequência das suas implantações aumentou com o crescimento do número de equipes de desenvolvimento. O número esperado definido de implantações é uma vez por semana. Você tem realizado implantações de forma regular e diariamente. Quando há um problema com o sistema de implantação e não é possível realizar as implantações, leva dias para que isso seja detectado.
- Antes, quando sua empresa oferecia suporte apenas durante o horário comercial, de segunda a sexta-feira. Você estabeleceu o próximo dia útil como a meta de tempo de resposta para incidentes. Recentemente, você iniciou a oferta de cobertura de suporte 24 horas por dia, 7 dias por semana, com uma meta de tempo de resposta de duas horas. Sua equipe noturna está sobrecarregada e os clientes estão insatisfeitos. Não há indicação de que haja problemas com os tempos de resposta a incidentes porque você está trabalhando com uma meta de próximo dia útil.

Benefícios do estabelecimento desta prática recomendada: Ao analisar e revisar os KPIs e as métricas, você entende como sua workload oferece suporte à obtenção dos resultados dos negócios e pode identificar onde é necessário melhorar para atingir suas metas de negócios.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Validar a obtenção de resultados e a eficácia dos KPIs e das métricas: crie uma visão de nível empresarial das atividades de operações para ajudá-lo a determinar se você está atendendo às

necessidades e a identificar áreas que precisam ser aprimoradas para atingir metas empresariais. Valide a eficácia dos KPIs e métricas e revise-os, se necessário.

- Uso de painéis do Amazon CloudWatch
- O que é análise de log?

Recursos

Documentos relacionados:

- Uso de painéis do Amazon CloudWatch
- O que é análise de log?

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.

Práticas recomendadas

- OPS10-BP01 Usar um processo para gerenciamento de eventos, incidentes e problemas
- OPS10-BP02 Ter um processo por alerta
- OPS10-BP03 Priorizar eventos operacionais com base no impacto nos negócios
- OPS10-BP04 Definir caminhos para escaladas
- OPS10-BP05 Habilitar notificações por push
- OPS10-BP06 Comunicar o status por meio de painéis
- OPS10-BP07 Automatizar respostas a eventos

OPS10-BP01 Usar um processo para gerenciamento de eventos, incidentes e problemas

Sua organização tem processos para lidar com eventos, incidentes e problemas. Eventos são coisas que ocorrem em sua workload que talvez não precisem de intervenção. Incidentes são eventos que requerem intervenção. Problemas são eventos recorrentes que exigem intervenção ou que não podem ser resolvidos. São necessários processos para reduzir o impacto desses eventos sobre os negócios e garantir respostas adequadas.

Quando incidentes e problemas acontecem em sua workload, você precisa de processos para lidar com eles. Como você vai comunicar o status do evento às partes interessadas? Quem supervisiona

e lidera a resposta? Quais são as ferramentas usadas para mitigar o evento? Esses são alguns exemplos de perguntas que você precisa responder para ter um processo de resposta sólido.

Os processos devem estar documentados em um local central e disponíveis a todos envolvidos com a workload. Se você não tiver uma wiki ou um armazenamento central de documentos, use um repositório de controle de versão. Você vai manter esses planos atualizados à medida que os processos evoluem.

Problemas são candidatos para automação. Esses eventos consomem o tempo que você poderia usar para inovar. Comece criando um processo repetível para mitigar o problema. Com o tempo, concentre-se na automação da mitigação ou correção do problema subjacente. Isso vai liberar tempo que você poderá dedicar ao desenvolvimento de melhorias para a workload.

Resultado desejado: sua organização tem processos para lidar com eventos, incidentes e problemas. Esses processos são documentados e armazenados em um local central. Eles são atualizados à medida que os processos mudam.

Antipadrões comuns:

- Um acidente ocorre durante um final de semana e o engenheiro de plantão não sabe o que fazer.
- Um cliente envia um e-mail informando que a aplicação está fora do ar. Você reinicializa o servidor para corrigir. Isso acontece com frequência.
- Há um incidente com várias equipes trabalhando de maneira independente para resolvê-lo.
- As implantações acontecem na workload sem serem registradas.

Benefícios do estabelecimento desta prática recomendada:

- Você tem uma trilha de auditoria de eventos na workload.
- O tempo para se recuperar de um incidente diminui.
- Os membros da equipe podem resolver incidentes e problemas de maneira consistente.
- Há um esforço mais consolidado na hora de investigar um incidente.

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

Orientação de implementação

Implementar essa prática recomendada significa que você está monitorando os eventos da workload. Você tem processos para lidar com incidentes e problemas. Os processos são documentados, compartilhados e atualizados com frequência. Problemas são identificados, priorizados e corrigidos.

Exemplo de cliente

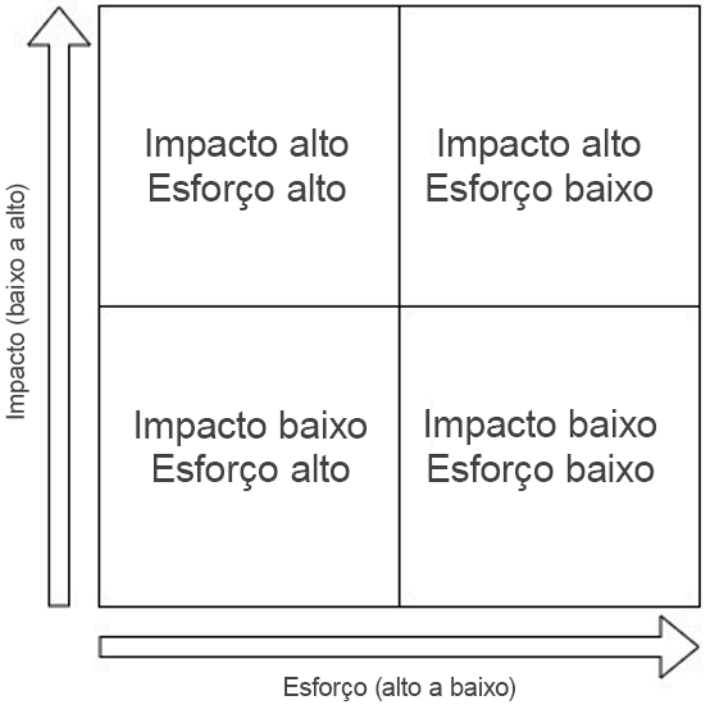
A AnyCompany Retail tem uma parte de sua wiki interna dedicada a processos para gerenciamento de eventos, incidentes e problemas. Todos os eventos são enviados para o Amazon EventBridge. Os problemas são identificados como OpsItems no OpsCenter do AWS Systems Manager e priorizados para correção, reduzindo a mão de obra não diferenciada. À medida que os processos mudam, eles são atualizados na wiki interna. Eles usam o AWS Systems Manager Incident Manager para gerenciar incidentes e coordenar os esforços de mitigação.

Etapas da implementação

1. Eventos
    - Monitore os eventos que acontecem na workload, mesmo que nenhuma intervenção humana seja necessária.
    - Trabalhe com as partes interessadas da workload para desenvolver uma lista de eventos que devem ser monitorados. Alguns exemplos são implantações concluídas ou aplicações de correções bem-sucedidas.
    - Você pode usar serviços como Amazon EventBridge ou Amazon Simple Notification Service para gerar eventos personalizados para monitoramento.
2. Incidentes
    - Comece definindo o plano de comunicação para incidentes. Quais partes interessadas devem ser informadas? Como você vai mantê-las informadas? Quem supervisiona os esforços de coordenação? Recomendamos a configuração de um canal de bate-papo interno para comunicação e coordenação.
    - Defina caminhos de encaminhamento para as equipes que oferecem suporte à workload, principalmente se a equipe não tiver uma rotação de plantão. Com base em seu nível de suporte, você também pode registrar um caso no AWS Support.
    - Crie um playbook para investigar o incidente. Isso deve incluir o plano de comunicação e etapas de investigação detalhadas. Inclua a verificação do AWS Health Dashboard na investigação.

- Documente seu plano de resposta a incidentes. Comunique o plano de gerenciamento de incidentes para que clientes internos e externos entendam as regras de engajamento e o que espera-se deles. Treine os membros de sua equipe sobre como usá-lo.
- Os clientes podem usar o Incident Manager para configurar e gerenciar seu respectivo plano de resposta a incidentes.
- Os clientes Enterprise Support podem solicitar o Workshop de gerenciamento de incidentes de seu gerente de conta técnico. Esse workshop guiado testa seu plano de resposta a incidentes e ajuda você a identificar áreas para melhoria.

3. Problemas
   - Os problemas devem ser identificados e monitorados em seu sistema de ITSM.
   - Identifique todos os problemas conhecidos e priorize-os em termos de esforço para corrigir e impacto na workload.



   - Resolva problemas de alto impacto e pouco esforço primeiro. Com esses resolvidos, passe para os problemas do quadrante de baixo impacto e pouco esforço.
   - Você pode usar o OpsCenter do Systems Manager para identificar esses problemas, anexar runbooks a eles e monitorá-los.

Nível de esforço do plano de implementação: médio. Você precisa de um processo e ferramentas para implementar essa prática recomendada. Documente seus processos e disponibilize-os para

todos que estão associados à workload. Atualize-os com frequência. Você tem um processo para gerenciar problemas e mitigá-los ou corrigi-los.

Recursos

Práticas recomendadas relacionadas:

- OPS07-BP03 Usar runbooks para realizar procedimentos: problemas conhecidos precisam de um runbook associado para que os esforços de mitigação sejam consistentes.
- OPS07-BP04 Usar manuais para investigar problemas: os incidentes precisam ser investigados usando playbooks.
- OPS11-BP02 Executar análise pós-incidente: sempre conduza uma autópsia depois de se recuperar de um incidente.

Documentos relacionados:

- Atlassian: gerenciamento de incidentes na era de DevOps
- Guia de resposta a incidentes de segurança da AWS
- Gerenciamento de incidentes na era de DevOps e SRE
- PagerDuty: o que é gerenciamento de incidentes?

Vídeos relacionados:

- AWS re:Invent 2020: Incident management in a distributed organization (AWS re:Invent 2020: gerenciamento de incidentes em uma organização distribuída)
- AWS re:Invent 2021 - Building next-gen applications with event-driven architectures (AWS re:Invent 2021 - criando aplicações de última geração com arquiteturas orientadas por eventos)
- AWS Supports You | Exploring the Incident Management Tabletop Exercise (AWS apoia você | Conhecendo a simulação teórica de gerenciamento de incidentes
- AWS Systems Manager Incident Manager - AWS Virtual Workshops (AWS Systems Manager Incident Manager - workshops virtuais da AWS)
- AWS What's Next ft. Incident Manager | AWS Events (Próximos passos na AWS com Incident Manager | Eventos da AWS)

Exemplos relacionados:

- workshop de ferramentas de gerenciamento e governança da AWS - OpsCenter
- Serviços proativos da AWS: workshop de gerenciamento de incidentes
- Como desenvolver uma aplicação orientada por eventos com o Amazon EventBridge
- Como desenvolver arquiteturas orientadas por eventos na AWS

Serviços relacionados:

- Amazon EventBridge
- Amazon SNS
- AWS Health Dashboard
- AWS Systems Manager Incident Manager
- OpsCenter do AWS Systems Manager

OPS10-BP02 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.

Antipadrões comuns:

- Seu sistema de monitoramento apresenta um stream de conexões aprovadas junto com outras mensagens. O volume de mensagens é tão grande que você perde mensagens de erro periódicas que exigem sua intervenção.
- Você recebe um alerta de que o site está inoperante. Não há um processo definido para quando isso acontece. Você é forçado a adotar uma abordagem ad hoc para diagnosticar e resolver o problema. Desenvolver esse processo conforme o uso estende o tempo para recuperação.

Benefícios do estabelecimento desta prática recomendada: Ao alertar somente quando uma ação é necessária, você impede que alertas de valor baixo ocultem alertas de valor alto. Ao ter um processo para alertas sempre acionáveis, você permite uma resposta consistente e imediata a eventos em seu ambiente.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Processo por alerta: qualquer evento para o qual você dispara um alerta deve ter uma resposta bem-definida (runbook ou manual) com um proprietário identificado especificamente (por exemplo, indivíduo, equipe ou função) responsável pela execução bem-sucedida. O desempenho da resposta pode ser automatizado ou conduzido por outra equipe, mas o proprietário é responsável por garantir que o processo ofereça os resultados esperados. Ao ter esses processos, você garante respostas eficazes e rápidas aos eventos de operações e pode impedir que eventos acionáveis sejam ocultados por notificações menos valiosas. Por exemplo, o auto scaling pode ser aplicado para dimensionar um front-end da web, mas a equipe de operações pode ser responsável por garantir que as regras e os limites de auto scaling sejam adequados para as necessidades de carga de trabalho.

Recursos

Documentos relacionados:

- Recursos do Amazon CloudWatch
- O que é o Amazon CloudWatch Events?

Vídeos relacionados:

- Build a monitoring plan

OPS10-BP03 Priorizar eventos operacionais com base no impacto nos negócios

Quando vários eventos demandarem intervenção, aborde primeiro os mais significativos para os negócios. Os impactos podem incluir perda de vida ou ferimentos, perda financeira ou danos à reputação ou confiança.

Antipadrões comuns:

- Você recebe uma solicitação de suporte para adicionar uma configuração de impressora para um usuário. Ao trabalhar no problema, você recebe uma solicitação de suporte informando que o site de varejo está inoperante. Depois de concluir a configuração da impressora para o usuário, você começa a trabalhar no problema do site.
- Você é notificado de que o site de varejo e o sistema de folha de pagamento estão inoperantes. Você não sabe para qual deve ter prioridade.

Benefícios do estabelecimento desta prática recomendada: A priorização de respostas aos incidentes com o maior impacto na empresa permite que você gerencie esse impacto.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Priorizar eventos operacionais com base no impacto empresarial: garanta que, quando vários eventos exigirem intervenção, aqueles que forem mais significativos para a empresa sejam abordados primeiro. Os impactos podem incluir perda de vida ou ferimentos, perda financeira, violações regulatórias ou danos à reputação ou à confiança.

OPS10-BP04 Definir caminhos para escaladas

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.

Saiba quando é necessária uma decisão humana antes que medidas sejam tomadas. Trabalhe com os tomadores de decisão para que essa decisão seja tomada antecipadamente e a ação seja pré-aprovada, para que a MTTR não seja estendida aguardando uma resposta.

Antipadrões comuns:

- Seu site de varejo está inoperante. Você não compreende o runbook para recuperar o site. Você começa a chamar colegas na expectativa de que alguém possa ajudá-lo.
- Você recebe um caso de suporte para um aplicativo inacessível. Você não tem permissões para administrar o sistema. Você não sabe quem tem. Você tenta entrar em contato com o proprietário do sistema que abriu o caso e não há resposta. Você não tem contatos do sistema e seus colegas não estão familiarizados com ele.

Benefícios do estabelecimento desta prática recomendada: Ao definir escalações, gatilhos para escalação e procedimentos para escalação, você permite a adição sistemática de recursos a um incidente a uma taxa apropriada para o impacto.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Definir caminhos para as escaladas: defina caminhos para as escaladas em seus runbooks e manuais, incluindo que é acionado pela escalada e os respectivos procedimentos. Por exemplo, escalação de um problema de engenheiros de suporte para engenheiros de suporte seniores quando a resolução do problema não estiver nos runbooks ou quando um período de tempo predefinido tiver decorrido. Outro exemplo de um caminho de escalação apropriado é dos engenheiros de suporte sênior à equipe de desenvolvimento para uma carga de trabalho quando os playbooks não conseguem identificar um caminho para a correção ou quando um período de tempo predefinido decorre. Identifique especificamente os proprietários de cada ação para garantir respostas eficazes e rápidas aos eventos de operações. Os escalonamentos podem incluir terceiros. Por exemplo, um provedor de conectividade de rede ou um fornecedor de software. Os escalonamentos podem incluir tomadores de decisão autorizados identificados para sistemas impactados.

OPS10-BP05 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.

Antipadrões comuns:

- Sua aplicação está sendo afetada por um incidente de negação de serviço distribuído e não responde há dias. Não há mensagem de erro. Você não enviou um e-mail de notificação. Você não enviou notificações por texto. Você não compartilhou informações nas mídias sociais. Seus clientes estão frustrados e procurando outros fornecedores que possam oferecer suporte a eles.
- Na segunda-feira, a aplicação teve problemas após a aplicação de um patch e ficou indisponível por algumas horas. Na terça-feira, a aplicação teve problemas após uma implantação de código e ficou inconfiável por algumas horas. Na quarta-feira, a aplicação teve problemas após uma implantação de código para mitigar uma vulnerabilidade de segurança associada ao patch com falha e ficou indisponível por algumas horas. Na quinta-feira, os frustrados clientes começaram a procurar outro fornecedor que lhes ofertasse suporte.
- Seu aplicativo ficará indisponível para manutenção neste fim de semana. Você não informa seus clientes. Alguns de seus clientes tinham atividades programadas que envolviam o uso do seu aplicativo. Eles ficam muito frustrados ao descobrir que seu aplicativo não está disponível.

Benefícios do estabelecimento desta prática recomendada: Ao definir notificações, gatilhos para notificações e procedimentos para notificações, você permite que o cliente seja informado e responda quando problemas com a carga de trabalho o afetarem.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Habilitar notificações por push: comunique-se diretamente com seus usuários (por e-mail ou por SMS, por exemplo) quando os serviços que eles usam forem afetados e quando os serviços retornarem às condições operacionais normais para permitir que os usuários tomem as medidas apropriadas.
  - Recursos do Amazon SES
  - O que é o Amazon SES?
  - Configurar as notificações do Amazon SNS

Recursos

Documentos relacionados:

- Recursos do Amazon SES
- Configurar as notificações do Amazon SNS
- O que é o Amazon SES?

OPS10-BP06 Comunicar o status por meio de painéis

Forneça painéis personalizados para seus públicos-alvo (por exemplo, equipes técnicas internas, liderança e clientes) para comunicar o status operacional atual dos negócios e fornecer métricas de interesse.

Você pode criar painéis usando o Painéis do Amazon CloudWatch em páginas de início personalizáveis no console do CloudWatch. Ao usar serviços de inteligência de negócios, como o Amazon QuickSight , você pode criar e publicar painéis interativos da carga de trabalho e da integridade operacional (por exemplo, taxas de pedidos, usuários conectados e tempos de transação). Crie painéis contendo visualizações em nível de sistema e de negócios de suas métricas.

Antipadrões comuns:

- Mediante solicitação, você executa um relatório sobre a utilização atual da aplicação para a gerência.
- Durante um incidente, você é contatado a cada vinte minutos por um proprietário do sistema preocupado, que deseja saber se ele já foi corrigido.

Benefícios do estabelecimento desta prática recomendada: Ao criar painéis, você permite o acesso por autoatendimento às informações, permitindo que os clientes se informem e determinem se precisam executar ações.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Comunicar o status por meio de painéis: forneça painéis personalizados para seus públicos-alvo (por exemplo, equipes técnicas internas, liderança e clientes) para comunicar o status operacional atual dos negócios e fornecer métricas de interesse. Fornecer uma opção de autoatendimento para informações de status reduz a interrupção das solicitações de status de campo pela equipe de operações. Os exemplos incluem os painéis do Amazon CloudWatch e o AWS Health Dashboard.
  - CloudWatch dashboards create and use customized metrics views (Os painéis do CloudWatch criam e usam visualizações de métricas personalizadas)

Recursos

Documentos relacionados:

- Amazon QuickSight
- CloudWatch dashboards create and use customized metrics views (Os painéis do CloudWatch criam e usam visualizações de métricas personalizadas)

OPS10-BP07 Automatizar respostas a eventos

Automatize as respostas aos eventos para reduzir erros causados por processos manuais e garantir respostas rápidas e consistentes.

Existem várias maneiras de automatizar a execução de ações de runbook e manual na AWS. Para responder a um evento de alteração de estado nos seus recursos da AWS, ou de seus próprios eventos personalizados, você deve criar regras do CloudWatch Events para acionar respostas

por meio de alvos do CloudWatch (por exemplo, funções do Lambda, tópicos do Amazon Simple Notification Service (Amazon SNS), tarefas do Amazon ECS e automação do AWS Systems Manager).

Para responder a uma métrica que ultrapassa um limite para um recurso (por exemplo, tempo de espera), você deve criar alarmes do CloudWatch para executar uma ou mais ações usando as ações do Amazon EC2, as ações do Auto Scaling ou enviar uma notificação para um tópico do Amazon SNS. Se for necessário executar ações personalizadas em resposta a um alarme, chame o Lambda por meio de uma notificação do Amazon SNS. Use o Amazon SNS para publicar notificações de eventos e mensagens de escalação para manter as pessoas informadas.

A AWS também é compatível com sistemas de terceiros por meio das APIs e SDKs de serviço da AWS. Existem várias ferramentas de monitoramento fornecidas por parceiros da AWS e por terceiros que permitem monitoramento, notificações e respostas. Algumas dessas ferramentas são New Relic, Splunk, Loggly, SumoLogic e Datadog.

Mantenha procedimentos manuais críticos disponíveis para uso quando houver falha em procedimentos automatizados.

Antipadrões comuns:

- Um desenvolvedor verifica seu código. Esse evento poderia ter sido usado para iniciar uma compilação e, em seguida, executar testes, mas, em vez disso, nada acontece.
- Sua aplicação registra um erro específico em log antes de parar de funcionar. O procedimento para reiniciar o aplicativo é bem compreendido e pode ter um script. Você pode usar o evento de log para invocar um script e reiniciar o aplicativo. Em vez disso, quando o erro acontece às 3 da manhã de domingo, você é despertado como o recurso de plantão responsável pela correção do sistema.

Benefícios do estabelecimento desta prática recomendada: Ao usar respostas automatizadas a eventos, você reduz o tempo de resposta e limita a introdução de erros oriundos de atividades manuais.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Automatizar respostas a eventos: automatize respostas a eventos para reduzir erros causados por processos manuais e garantir respostas rápidas e consistentes.

- O que é o Amazon CloudWatch Events?
- Criação de uma regra do CloudWatch Events que aciona um evento
- Criação de uma regra do CloudWatch Events que aciona uma chamada de API da AWS usando o AWS CloudTrail
- Exemplos de eventos do CloudWatch Events de serviços compatíveis

Recursos

Documentos relacionados:

- Recursos do Amazon CloudWatch
- Exemplos de eventos do CloudWatch Events de serviços compatíveis
- Criação de uma regra do CloudWatch Events que aciona uma chamada de API da AWS usando o AWS CloudTrail
- Criação de uma regra do CloudWatch Events que aciona um evento
- O que é o Amazon CloudWatch Events?

Vídeos relacionados:

- Build a monitoring plan

Exemplos relacionados:

Evoluir

Pergunta

- OPS 11 Como você faz para que as operações evoluam?

OPS 11 Como você faz para que as operações evoluam?

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

Práticas recomendadas

- OPS11-BP01 Ter um processo para a melhoria contínua

- OPS11-BP02 Executar análise pós-incidente
- OPS11-BP03 Implementar loops de feedback
- OPS11-BP04 Executar o gerenciamento de conhecimento
- OPS11-BP05 Definir motivadores de melhoria
- OPS11-BP06 Validar insights
- OPS11-BP07 Fazer análises das métricas de operações
- OPS11-BP08 Documentar e compartilhar as lições aprendidas
- OPS11-BP09 Alocar tempo para fazer melhorias

OPS11-BP01 Ter um processo para a melhoria contínua

Avalie e priorize regularmente oportunidades de melhorias para concentrar os esforços onde eles possam oferecer os maiores benefícios.

Antipadrões comuns:

- Você documentou os procedimentos necessários para criar um ambiente de desenvolvimento ou teste. Você pode usar o CloudFormation para automatizar o processo, mas, em vez disso, faz isso manualmente no console.
- Os testes mostram que a grande maioria da utilização da CPU dentro do aplicativo está em um pequeno conjunto de funções ineficientes. Você pode se concentrar em melhorá-las e reduzir seus custos, mas foi encarregado de criar um novo recurso de usabilidade.

Benefícios do estabelecimento desta prática recomendada: A melhoria contínua oferece um mecanismo para avaliar regularmente oportunidades de melhoria, priorizar oportunidades e concentrar esforços onde eles possam fornecer os maiores benefícios.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Definir processos para a melhoria contínua: avalie e priorize regularmente as oportunidades de melhoria para concentrar os esforços naquilo que possa fornecer maiores benefícios. Implemente mudanças para melhorar e avaliar os resultados para determinar o sucesso. Se os resultados não satisfizerem as metas e a melhoria ainda for uma prioridade, itere usando cursos de ação alternativos. Seus processos operacionais devem incluir tempo e recursos dedicados para possibilitar melhorias incrementais contínuas.

OPS11-BP02 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.

Antipadrões comuns:

- Você administra um servidor de aplicativos. Aproximadamente a cada 23 horas e 55 minutos, todas as sessões ativas são encerradas. Você tentou identificar o que está errado no servidor de aplicativos. Você suspeita que possa ser um problema de rede, mas não consegue obter colaboração da equipe da rede, pois ela está muito ocupada para ajudar você. Você não tem um processo predefinido a seguir para obter suporte e coletar as informações necessárias para determinar o que está acontecendo.
- Você teve perda de dados em sua carga de trabalho. Esta é a primeira vez que isso acontece e a causa não é óbvia. Você decide que não é importante porque pode recriar os dados. A perda de dados começa a ocorrer com maior frequência, afetando seus clientes. Isso também coloca uma sobrecarga operacional adicional à medida que você restaura os dados ausentes.

Benefícios do estabelecimento desta prática recomendada: Ter processos predefinidos para determinar componentes, condições, ações e eventos que contribuíram para um incidente permite identificar oportunidades de melhoria.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

- Usar um processo para determinar os fatores contribuintes: analise todos os incidentes que impactam os clientes. Tenha um processo para identificar e documentar as causas de um incidente para que você possa desenvolver atenuações para limitar ou impedir a recorrência e para desenvolver procedimentos para respostas rápidas e eficazes. Se for apropriado, comunique as causas de forma direcionada para o público-alvo.

OPS11-BP03 Implementar loops de feedback

Os loops de feedback fornecem insights que levem a ações concretas e orientem a tomada de decisões. Crie loops de feedback em seus procedimentos e workloads. Isso ajuda a identificar

problemas e áreas que precisam de melhorias. Eles também validam os investimentos feitos em melhorias. Esses loops de feedback são a base para melhorar continuamente sua workload.

Os loops de feedback se enquadram em duas categorias: feedback imediato e análise retrospectiva. O feedback imediato é coletado por meio da avaliação do desempenho e dos resultados das atividades de operações. Esse feedback vem de membros da equipe, clientes ou do resultado automático da atividade. O feedback imediato é recebido de elementos como testes A/B e do envio de novos recursos, e é essencial para antecipar-se à falha.

A análise retrospectiva é realizada regularmente para obter feedback da avaliação de resultados e métricas operacionais ao longo do tempo. Essa retrospectiva ocorre ao final de um sprint, com certa frequência ou após grandes lançamentos ou eventos. Esse tipo de loop de feedback valida investimentos em operações ou na workload. Ele ajuda a medir o sucesso e valida sua estratégia.

Resultado desejado: o feedback imediato e a análise retrospectiva são usados para promover melhorias. Há um mecanismo para obter o feedback de usuários e membros da equipe. A análise retrospectiva é usada para identificar tendências que promovam melhorias.

Antipadrões comuns:

- Você lança um recurso, mas não há uma maneira de receber feedback de clientes sobre ele.
- Depois de investir em melhorias de operações, você não realiza uma retrospectiva para validá-las.
- Você coleta feedback dos clientes, mas não os avalia regularmente.
- Os loops de feedback levam a itens de ação propostos, mas não estão incluídos no processo de desenvolvimento de software.
- Os clientes não recebem feedback sobre as melhorias que propuseram.

Benefícios de estabelecer esta prática recomendada:

- Você pode trabalhar partindo do feedback do cliente para gerar outros recursos.
- A cultura da sua organização pode reagir às mudanças mais rapidamente.
- As tendências são usadas para identificar oportunidades de melhoria.
- As retrospectivas validam os investimentos feitos na workload e nas operações.

Nível de risco exposto se essa prática recomendada não for estabelecida: alto

Orientação para implementação

A implementação dessa prática recomendada significa que você usa tanto o feedback imediato como a análise de retrospectiva. Esses loops de feedback geram melhorias. Há muitos mecanismos para o feedback imediato, incluindo pesquisas, enquetes com clientes ou formulários de feedback. Sua organização também pode usar as retrospectivas para identificar oportunidades de melhoria e validar iniciativas.

Exemplo de cliente

A Loja UmaEmpresa criou um formulário online pelo qual os clientes podem dar feedback ou relatar problemas. Durante as reuniões semanais, o feedback dos usuários é avaliado pela equipe de desenvolvimento de software. O feedback é usado regularmente para conduzir a evolução da plataforma. É feita uma retrospectiva ao final de cada sprint para identificar itens que eles desejam melhorar.

Etapas da implementação

1. Feedback imediato
   - Você precisa de um mecanismo para receber feedback de clientes e membros da equipe. Suas atividades de operações também podem ser configuradas para oferecer feedback automático.
   - Sua organização precisa de um processo para avaliar esse feedback, determinar o que precisa ser melhorado e programar a melhoria.
   - O feedback deve ser adicionado ao seu processo de desenvolvimento de software.
   - À medida que você faz melhorias, dê um retorno a quem enviou o feedback.
     - Você pode usar o AWS Systems Manager OpsCenter para criar e monitorar essas melhorias como OpsItems.
2. Análise retrospectiva
   - Faça retrospectivas ao final de um ciclo de desenvolvimento, com certa frequência ou após um grande lançamento.
   - Faça uma reunião de retrospectiva com as partes interessadas envolvidas na workload.
   - Crie três colunas em um quadro branco ou uma planilha: “Parar”, “Iniciar” e “Manter”.
     - A coluna “Parar” é para o que você deseja que a equipe pare de fazer.
     - A coluna “Iniciar” é para ideias que você deseja começar a fazer.
     - A coluna “Manter” é para os itens que você deseja continuar fazendo.
   - Caminhe pela sala e colete o feedback das partes interessadas.

- Priorize o feedback. Atribua ações e partes interessadas aos itens de “Iniciar” e “Manter”.
- Adicione as ações ao processo de desenvolvimento de software e comunique as atualizações de status às partes interessadas à medida que as melhorias são implementadas.

Nível de esforço do plano de implementação: médio. Para implementar essa prática recomendada, você precisa de uma maneira para receber feedback imediato e analisá-lo. Além disso, você precisa estabelecer um processo de análise de retrospectiva.

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP01 Avaliar as necessidades dos clientes externos: loops de feedback são um mecanismo para coletar as necessidades de clientes externos.
- OPS01-BP02 Avalie as necessidades dos clientes internos: as partes interessadas internas podem usar loops de feedback para comunicar necessidades e requisitos.
- OPS11-BP02 Executar análise pós-incidente: a análise pós-incidente é uma forma importante de análise retrospectiva conduzida após os incidentes.
- OPS11-BP07 Fazer análises das métricas de operações: as avaliações das métricas de operações identificam tendências e áreas para melhorias.

Documentos relacionados:

- 7 Pitfalls to Avoid When Building a CCOE (Sete obstáculos a evitar ao criar um CCoE)
- Atlassian Team Playbook - Retrospectives (Manual da equipe do Atlassian: retrospectivas)
- Email Definitions: Feedback Loops (Definições de e-mail: loops de feedback)
- Establishing Feedback Loops Based on the AWS Well-Architected Framework Review (Como estabelecer loops de feedback com base na avaliação do AWS Well-Architected Framework)
- IBM Garage Methodology - Hold a retrospective (Metodologia IBM Garage: faça uma retrospectiva)
- Investopedia – The PDCS Cycle (Investopédia: o ciclo de PDCS)
- Maximizing Developer Effectiveness by Tim Cochran (Como maximizar a eficácia do desenvolvedor, por Tim Cochran)
- Operations Readiness Reviews (ORR) Whitepaper - Iteration (Whitepaper de análises de preparação de operações (ORR): iteração)
- TIL CSI - Continual Service Improvement (CSI de TIL: melhoria de serviço contínua)

- When Toyota met e-commerce: Lean at Amazon (Quando a Toyota chegou ao comércio eletrônico: confiança na Amazon)

Vídeos relacionados:

- Building Effective Customer Feedback Loops (Como criar loops de feedback eficazes de clientes)

Exemplos relacionados:

- Astuto - Open source customer feedback tool (Astuto: ferramenta de código aberto de feedback de clientes)
- AWS Solutions - QnABot on AWS (Soluções da AWS: QnABot na AWS)
- Fider - A platform to organize customer feedback (Fider: uma plataforma para organizar feedback de clientes)

Serviços relacionados:

- AWS Systems Manager OpsCenter

OPS11-BP04 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.

Antipadrões comuns:

- Um único cliente frustrado abre um caso de suporte para uma nova solicitação de recurso de produto para resolver um problema percebido. Ele é adicionado à lista de melhorias de prioridade.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Executar gerenciamento do conhecimento: verifique se 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. Mantenha mecanismos para identificar o conteúdo necessário, o conteúdo que precisa de atualização e o conteúdo que deve ser arquivado para que ele não seja mais referenciado.

OPS11-BP05 Definir motivadores de melhoria

Identifique os condutores de melhoria para ajudá-lo a avaliar e priorizar as oportunidades.

Na AWS, é possível agregar os logs de todas as suas atividades operacionais, workloads e infraestrutura para criar um histórico de atividades detalhado. Assim, é possível usar as ferramentas da AWS para analisar as operações e a integridade da workload ao longo do tempo (por exemplo, identificar tendências, correlacionar eventos e atividades a resultados e comparar e contrastar ambientes e sistemas) para revelar oportunidades de melhoria com base em seus motivadores.

Use o CloudTrail para rastrear a atividade da API (por meio do AWS Management Console, da CLI, de SDKs e de APIs) para saber o que está acontecendo nas suas contas. Rastreie as atividades de implantação das ferramentas do desenvolvedor da AWS com o CloudTrail e o CloudWatch. Isso adicionará um histórico detalhado das atividades de suas implantações e seus resultados aos dados de logs do CloudWatch Logs.

Exporte seus dados de log para o Amazon S3 para armazenamento de longo prazo. Com o uso do AWS Glue, você descobre e prepara seus dados de log no Amazon S3 para análise. Uso Amazon Athena, por meio de sua integração nativa com o AWS Glue, para analisar os dados de logs. Use uma ferramenta de business intelligence, como o Amazon QuickSight , para visualizar, explorar e analisar os dados.

Antipadrões comuns:

- Você tem um script que funciona, mas não é elegante. Você investe tempo para reescrevê-lo. Agora, ele é uma obra de arte.
- Sua startup está tentando obter outro conjunto de financiamento de um capitalista de risco. Ele quer que você demonstre conformidade com o PCI DSS. Você quer deixá-lo contente, então documenta sua conformidade e perde uma data de entrega para um cliente, perdendo esse cliente. Não foi algo errado, mas agora você se pergunta se foi o certo a se fazer.

Benefícios do estabelecimento desta prática recomendada: Ao determinar os critérios que você deseja implantar para melhorar, é possível minimizar o impacto das motivações baseadas em eventos ou investimentos emocionais.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Compreender as motivações para melhoria: só faça alterações em um sistema quando o resultado desejado for suportado.
  - Capacidades desejadas: avalie as capacidades e os recursos desejados ao avaliar oportunidades de melhoria.
    - Novidades da AWS
  - Problemas inaceitáveis: avalie problemas, bugs e vulnerabilidades inaceitáveis ao avaliar oportunidades de melhoria.
    - Boletins de segurança mais recentes da AWS
    - AWS Trusted Advisor
  - Requisitos de conformidade: avalie as atualizações e alterações necessárias para manter a conformidade com a regulamentação e com a política, ou para permanecer sob o suporte de terceiros ao analisar as oportunidades de melhoria.
    - Conformidade da AWS
    - Programas de conformidade da AWS
    - Notícias recentes sobre conformidade da AWS

Recursos

Documentos relacionados:

- Amazon Athena
- Amazon QuickSight
- Conformidade da AWS
- Notícias recentes sobre conformidade da AWS
- Programas de conformidade da AWS
- AWS Glue
- Boletins de segurança mais recentes da AWS
- AWS Trusted Advisor
- Exporte seus dados de log para o Amazon S3
- Novidades da AWS

OPS11-BP06 Validar 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.

Antipadrões comuns:

- Você vê que a utilização da CPU está em 95% em um sistema e prioriza encontrar uma maneira de reduzir a carga no sistema. Você determina que a melhor ação é expandir. O sistema é um transcodificador e foi dimensionado para ser executado com 95% de utilização da CPU o tempo todo. O proprietário do sistema poderia ter explicado a situação se você tivesse entrado em contato com ele. Seu tempo foi perdido.
- Um proprietário do sistema sustenta que o sistema é de missão crítica. O sistema não foi colocado em um ambiente de alta segurança. Para melhorar a segurança, você implementa os controles de detecção e prevenção adicionais necessários para sistemas de missão crítica. Você notifica o proprietário do sistema de que o trabalho foi concluído e que ele será cobrado pelos recursos adicionais. Na discussão após essa notificação, o proprietário do sistema aprende que há uma definição formal para sistemas de missão crítica que o sistema dele não atende.

Benefícios do estabelecimento desta prática recomendada: Ao validar insights com proprietários de empresas e especialistas no assunto, você pode estabelecer um entendimento comum e orientar de maneira mais eficaz a melhoria.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Validar insights: envolva-se com proprietários de empresas e especialistas no assunto para garantir que haja entendimento e concordância comuns sobre o significado dos dados coletados. Identifique preocupações adicionais, possíveis impactos e determine as ações.

OPS11-BP07 Fazer análises das 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.

Procure oportunidades para melhorar em todos os seus ambientes (por exemplo, desenvolvimento, teste e produção).

Antipadrões comuns:

- Houve uma promoção de varejo significativa que foi interrompida por sua janela de manutenção. A empresa continua sem saber que existe uma janela de manutenção padrão que pode ser atrasada se houver outros eventos que afetam os negócios.
- Você sofreu uma interrupção prolongada devido ao uso de uma biblioteca com bugs geralmente utilizada em sua organização. Desde então, você migrou para uma biblioteca confiável. As outras equipes da organização não sabem que estão em risco. Se você se reunisse regularmente e analisasse esse incidente, eles ficariam conscientes do risco.
- A performance do transcodificador tem diminuído constantemente e está afetando a equipe de mídia. Ainda não é algo terrível. Você não terá a oportunidade de descobrir até que seja ruim o suficiente para causar um incidente. Se você analisasse as métricas de operações com a equipe de mídia, haveria uma oportunidade para que a mudança nas métricas e a experiência deles fossem reconhecidas e o problema fosse resolvido.
- Você não está analisando a satisfação dos SLAs do cliente. Você está tendendo a não cumprir os SLAs de seus clientes. Há penalidades financeiras relacionadas ao não cumprimento de SLAs dos clientes. Se você se reunisse regularmente para analisar as métricas desses SLAs, teria a oportunidade de reconhecer e resolver o problema.

Benefícios do estabelecimento desta prática recomendada: Ao realizar reuniões regularmente para analisar métricas, eventos e incidentes de operações, você mantém um entendimento comum entre as equipes, compartilha as lições aprendidas e pode priorizar e direcionar melhorias.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Análises das métricas das operações: execute regularmente análises retrospectivas das métricas de operações com participantes de equipes de diferentes áreas do negócio. Envolva as partes interessadas, incluindo as equipes de negócios, desenvolvimento e operações, para validar suas descobertas de feedback imediato e análise retrospectiva e para compartilhar as lições aprendidas. Use suas ideias para identificar oportunidades de melhoria e possíveis cursos de ação.
  - Amazon CloudWatch

- Uso de métricas do Amazon CloudWatch
- Publicar métricas personalizadas
- Referência de métricas e de dimensões do Amazon CloudWatch

Recursos

Documentos relacionados:

- Amazon CloudWatch
- Referência de métricas e de dimensões do Amazon CloudWatch
- Publicar métricas personalizadas
- Uso de métricas do Amazon CloudWatch

OPS11-BP08 Documentar e compartilhar as lições aprendidas

Documente e compartilhe as lições aprendidas das atividades operacionais, para que possa usá-las internamente e entre equipes.

Você deve compartilhar o que suas equipes aprendem para aumentar os benefícios em toda a organização. Você desejará compartilhar informações e recursos para evitar erros que podem ser evitados e facilitar os esforços de desenvolvimento. isso permitirá que você se concentre no fornecimento dos recursos desejados.

Use o AWS Identity and Access Management (IAM) para definir permissões que permitem acesso controlado aos recursos que você deseja compartilhar dentro e entre contas. Você deve usar os repositórios do AWS CodeCommit com controle de versão para compartilhar bibliotecas de aplicativos, procedimentos com script, documentações de procedimentos e outras documentações do sistema. Compartilhe seus padrões de computação partilhando o acesso às suas AMIs e autorizando o uso de suas funções do Lambda entre contas. Você também deve compartilhar seus padrões de infraestrutura como modelos do AWS CloudFormation.

Por meio de APIs e SDKs da AWS, é possível integrar ferramentas e repositórios externos e de terceiros (por exemplo, GitHub, BitBucket e SourceForge). Ao compartilhar o que você aprendeu e desenvolveu, tenha cuidado para estruturar as permissões para garantir a integridade dos repositórios compartilhados.

Antipadrões comuns:

- Você sofreu uma interrupção prolongada devido ao uso de uma biblioteca com bugs geralmente utilizada em sua organização. Desde então, você migrou para uma biblioteca confiável. As outras equipes em sua organização não sabem que estão em risco. Se você documentasse e compartilhasse sua experiência com essa biblioteca, eles ficariam cientes do risco.
- Você identificou um caso de borda em um microsserviço compartilhado internamente que causa a queda das sessões. Você atualizou suas chamadas para o serviço para evitar esse caso extremo. As outras equipes da organização não sabem que estão em risco. Se você documentasse e compartilhasse sua experiência com essa biblioteca, eles ficariam cientes do risco.
- Você encontrou uma maneira de reduzir significativamente os requisitos de utilização da CPU para um dos seus microsserviços. Você não sabe se alguma outra equipe poderia aproveitar essa técnica. Se você documentasse e compartilhasse sua experiência com essa biblioteca, eles teriam a oportunidade de aproveitá-la.

Benefícios do estabelecimento desta prática recomendada: Compartilhe as lições aprendidas para apoiar melhorias e maximizar os benefícios da experiência.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Documentar e compartilhar as lições aprendidas: tenha procedimentos para documentar as lições aprendidas com a execução de atividades operacionais e análises retrospectivas, para que possam ser usadas por outras equipes.
  - Compartilhar lições aprendidas: tenha procedimentos para compartilhar as lições aprendidas e os artefatos associados entre as equipes. Por exemplo, compartilhe procedimentos atualizados, orientações, governança e práticas recomendadas por meio de um wiki acessível. Compartilhe scripts, códigos e bibliotecas por meio de um repositório comum.
    - Delegação de acesso ao ambiente da AWS
    - Compartilhar um repositório do AWS CodeCommit
    - Fácil autorização das funções do AWS Lambda
    - Compartilhamento de uma AMI com contas específicas da AWS
    - Acelerar o compartilhamento de modelos com uma URL do designer do AWS CloudFormation
    - Usar o AWS Lambda com o Amazon SNS

Recursos

Documentos relacionados:

- Fácil autorização das funções do AWS Lambda
- Compartilhar um repositório do AWS CodeCommit
- Compartilhamento de uma AMI com contas específicas da AWS
- Acelerar o compartilhamento de modelos com uma URL do designer do AWS CloudFormation
- Usar o AWS Lambda com o Amazon SNS

Vídeos relacionados:

- Delegação de acesso ao ambiente da AWS

OPS11-BP09 Alocar tempo para fazer melhorias

Dedique tempo e recursos em seus processos para possibilitar melhorias incrementais contínuas.

Na AWS, é possível criar duplicatas temporárias de ambientes, reduzindo o risco, o esforço e o custo da experimentação e dos testes. Esses ambientes duplicados podem ser usados para testar as conclusões de sua análise, experimentar e desenvolver e testar as melhorias planejadas.

Antipadrões comuns:

- Há um problema de performance conhecido no servidor de aplicativos. Ele é adicionado ao backlog por trás de cada implementação de recurso planejada. Se a taxa de adição de recursos planejados permanecer constante, o problema de performance nunca será resolvido.
- Para oferecer suporte à melhoria contínua, você aprova administradores e desenvolvedores usando todo o tempo extra para selecionar e implementar melhorias. Nenhuma melhoria é concluída.

Benefícios do estabelecimento desta prática recomendada: Ao dedicar tempo e recursos em seus processos, você possibilita melhorias incrementais contínuas.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Alocar tempo para fazer melhorias: dedique tempo e recursos em seus processos para possibilitar melhorias incrementais contínuas. Implemente alterações para melhorar e avaliar os resultados para determinar o sucesso. Se os resultados não satisfizerem as metas e a melhoria ainda for uma prioridade, siga cursos de ação alternativos.

Segurança

Tópicos

Fundamentos de segurança

Pergunta

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.

Práticas recomendadas

- SEC01-BP01 Separar as workloads usando contas
- SEC01-BP02 Proteger a Conta da AWS

- SEC01-BP03 Identificar e validar objetivos de controle
- SEC01-BP04 Manter-se atualizado sobre as ameaças à segurança
- SEC01-BP05 Manter-se atualizado com as recomendações de segurança
- SEC01-BP06 Automatizar testes e validação de controles de segurança em pipelines
- SEC01-BP07 Identificar e priorizar riscos usando um modelo de ameaça
- SEC01-BP08 Avaliar e implementar regularmente novos serviços e recursos de segurança

SEC01-BP01 Separar as workloads usando contas

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. Essa abordagem fornece limites e controles entre cargas de trabalho. A separação no nível da conta é altamente recomendada para isolar ambientes de produção de ambientes de desenvolvimento e teste, ou para determinar um limite lógico forte entre cargas de trabalho que processam dados de diferentes níveis de confidencialidade, conforme definido por requisitos de conformidade externos (como PCI-DSS ou HIPAA) e cargas de trabalho que não processam.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Usar o AWS Organizations: use o AWS Organizations para aplicar centralmente o gerenciamento baseado em políticas para várias Contas da AWS.
  - Conceitos básicos do AWS Organizations
  - How to use service control policies to set permission guardrails across accounts in your AWS Organization (Como usar políticas de controle de serviços para definir barreiras de proteção de permissão entre contas no AWS Organization)
- Considerar o AWS Control Tower: o AWS Control Tower oferece uma maneira fácil de configurar e controlar um novo ambiente seguro e multicontas da AWS com base nas práticas recomendadas.
  - AWS Control Tower

Recursos

Documentos relacionados:

- Práticas recomendadas do IAM

- Boletins de segurança
- AWS Security Audit Guidelines (Diretrizes de auditoria de segurança da AWS)

Vídeos relacionados:

- Como gerenciar ambientes da AWS de várias contas usando o AWS Organizations
- Security Best Practices the Well-Architected Way
- Usar o AWS Control Tower para administrar ambientes da AWS de várias contas

SEC01-BP02 Proteger a Conta da AWS

Há uma série de aspectos para proteger suas Contas da AWS, incluindo a proteger e não utilizar o usuário raize manter as informações de contato atualizadas. Você pode usar o AWS Organizations para gerenciar e controlar centralmente suas contas à medida que expande e dimensiona suas workloads na AWS. O AWS Organizations ajuda você a gerenciar contas, definir controles e configurar serviços em todas as suas contas.

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

Orientação de implementação

- Usar o AWS Organizations: use o AWS Organizations para aplicar centralmente o gerenciamento baseado em políticas para várias Contas da AWS.
  - Conceitos básicos do AWS Organizations
  - Como usar políticas de controle de serviço para definir barreiras de proteção de permissão entre contas no AWS Organization
- Limitar o uso do usuário raiz da AWS: somente use o usuário raiz para executar tarefas que o exijam especificamente.
  - Tarefas da AWS que exigem credenciais do usuário raiz da conta da AWS
- Habilitar a autenticação multifator (MFA) para o usuário raiz: habilite a MFA no usuário raiz da Conta da AWS, se o AWS Organizations não estiver gerenciando usuários raiz para você.
  - Usuário raiz
- Altere periodicamente a senha do usuário raiz. Alterar a senha do usuário raiz reduz o risco de que uma senha salva possa ser usada. Isso é particularmente importante se você não estiver usando o AWS Organizations e alguém tiver acesso físico.
  - Alteração da senha do usuário raiz da Conta da AWS

- Habilite a notificação quando o usuário raiz da Conta da AWS for usado: ser notificado automaticamente reduz o risco.
  - Como receber notificações quando suas chaves de acesso raiz da Conta da AWS são usadas
- Restringir o acesso a regiões adicionadas recentemente: para novas regiões da Regiões da AWS, os recursos do IAM, como usuários e perfis, serão propagados somente para as regiões habilitadas.
  - Configuração das permissões para habilitar contas para as próximas Regiões da AWS
- Considere o AWS CloudFormation StackSets: o CloudFormation StackSets pode ser usado para implantar recursos, incluindo políticas, perfis e grupos do IAM, em diferentes regiões e Contas da AWS por meio de um modelo aprovado.
  - Use o CloudFormation StackSets

Recursos

Documentos relacionados:

- AWS Control Tower
- AWS Security Audit Guidelines (Diretrizes de auditoria de segurança da AWS)
- Práticas recomendadas do IAM
- Boletins de segurança

Vídeos relacionados:

- Enable AWS adoption at scale with automation and governance (Habilite a adoção da AWS em escala com automação e governança)
- Security Best Practices the Well-Architected Way

Exemplos relacionados:

- Laboratório: usuário raiz e Conta da AWS

SEC01-BP03 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.

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

Orientação de implementação

- Identificar requisitos de conformidade: descubra os requisitos organizacionais, legais e de conformidade que a sua workload precisa cumprir.
- Identificar recursos de conformidade da AWS: identifique os recursos da AWS disponíveis para ajudar você com a conformidade.
  - https://aws.amazon.com/compliance/
  - https://aws.amazon.com/artifact/

Recursos

Documentos relacionados:

- AWS Security Audit Guidelines (Diretrizes de auditoria de segurança da AWS)
- Boletins de segurança

Vídeos relacionados:

- AWS Security Hub: Manage Security Alerts and Automate Compliance (AWS Security Hub: gerenciamento de alertas de segurança e automatização da governança)
- Security Best Practices the Well-Architected Way

SEC01-BP04 Manter-se atualizado sobre as ameaças à segurança

Para ajudar a definir e implementar os controles apropriados, reconheça vetores de ataque mantendo-se a par das ameaças de segurança mais recentes. Consuma o AWS Managed Services para facilitar o recebimento de notificações de comportamentos inesperados ou incomuns em suas contas da AWS. Investigue usando ferramentas de parceiros da AWS ou feeds de informações sobre ameaças de terceiros como parte de seu fluxo de informações de segurança. A lista de vulnerabilidades e exposições comuns (CVEs) contém vulnerabilidades de segurança cibernética divulgadas publicamente que você pode usar para se manter atualizado.

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

Orientação de implementação

- Inscreva-se em fontes de inteligência de ameaças: analise regularmente as informações de inteligência de ameaças de várias fontes relevantes sobre as tecnologias usadas na sua workload.
  - Lista de vulnerabilidades e exposições comuns
- Considerar AWS Shield Advanced : oferece visibilidade quase em tempo real das fontes de inteligência, se sua workload for acessível pela Internet.

Recursos

Documentos relacionados:

- AWS Security Audit Guidelines (Diretrizes de auditoria de segurança da AWS)
- AWS Shield
- Boletins de segurança

Vídeos relacionados:

- Security Best Practices the Well-Architected Way

SEC01-BP05 Manter-se atualizado com as recomendações de segurança

Mantenha-se atualizado com as recomendações de segurança da AWS e do setor para evoluir a postura de segurança de sua workload. Boletins de segurança da AWS contêm informações importantes sobre notificações de segurança e privacidade.

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

Orientação de implementação

- Siga as atualizações da AWS: inscreva-se ou verifique regularmente novas recomendações e dicas.
  - Laboratórios do AWS Well-Architected
  - Blog de segurança da AWS
  - Documentação do serviço da AWS
- Inscreva-se para receber as novidades do setor: consulte regularmente os feeds de notícias de várias fontes relevantes às tecnologias usadas na sua workload.

- Exemplo: lista de vulnerabilidade e exposições comuns

Recursos

Documentos relacionados:

- Boletins de segurança

Vídeos relacionados:

- Security Best Practices the Well-Architected Way

SEC01-BP06 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 modelos de infraestrutura como código, para detectar vulnerabilidades de segurança, irregularidades e desvios da uma linha de base estabelecida em cada estágio. O AWS CloudFormation Guard pode ajudar você a verificar se os modelos do CloudFormation são seguros, economizar tempo e reduzir o risco de erro de configuração.

É fundamental reduzir o número de configurações incorretas de segurança introduzidas em um ambiente de produção. Portanto, quanto mais você puder controlar a qualidade e reduzir os defeitos no processo de construção, melhor. Projete pipelines de integração e implantação contínua (CI/CD) para testar problemas de segurança sempre que possível. Os pipelines de CI/CD oferecem a oportunidade de aumentar a segurança em cada estágio de criação e entrega. As ferramentas de segurança de CI/CD também devem estar sempre atualizadas para mitigar as ameaças em constante evolução.

Acompanhe as alterações na configuração de workload para ajudar na auditoria de conformidade, gerenciamento de alterações e investigações que possam ser aplicáveis. Você pode usar o AWS Config para registrar e avaliar seus recursos da AWS e de terceiros. Ele permite auditar e avaliar continuamente a conformidade geral com regras e pacotes de conformidade, que são coleções de regras com ações de correção.

O rastreamento de alterações deve incluir alterações planejadas, que fazem parte do processo de controle de alterações da sua organização [às vezes chamado de MACD, de Move, Add, Change,

Delete (Mover, Adicionar, Alterar, Excluir)], alterações não planejadas e alterações inesperadas, como incidentes. Podem ocorrer alterações na infraestrutura, mas também podem estar relacionadas a outras categorias, como alterações em repositórios de código, imagens de máquina e alterações de inventário de aplicações, alterações de processos e políticas ou alterações de documentação.

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

Orientação de implementação

- Automatize o gerenciamento de configuração: aplique e valide configurações seguras automaticamente usando uma ferramenta ou um serviço de gerenciamento de configuração.
  - AWS Systems Manager
  - AWS CloudFormation
  - Configurar um pipeline CI/CD na AWS

Recursos

Documentos relacionados:

- Como usar políticas de controle de serviço para definir barreiras de proteção de permissão entre contas no AWS Organization

Vídeos relacionados:

- Como gerenciar ambientes da AWS de várias contas usando o AWS Organizations
- Security Best Practices the Well-Architected Way

SEC01-BP07 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 ameaças e adapte os controles de segurança para prevenir, detectar e responder. Revise e mantenha essas informações no contexto do cenário de segurança em evolução.

A modelagem de ameaças fornece uma abordagem sistemática para ajudar a encontrar e resolver problemas de segurança no início do processo de design. Quanto mais cedo melhor, pois as mitigações têm um custo mais baixo em comparação com o final do ciclo de vida.

As etapas principais típicas do processo de modelagem de ameaças são:

1. Identificar ativos, atores, pontos de entrada, componentes, casos de uso e níveis de confiança e incluí-los em um diagrama de design.
2. Identificar uma lista de ameaças.
3. Para cada ameaça, identifique mitigações, que podem incluir implementações de controle de segurança.
4. Criar e revisar uma matriz de risco para determinar se a ameaça foi mitigada de forma adequada.

A modelagem de ameaças é mais eficaz quando feita no nível da workload (ou recurso de workload), garantindo que todo o contexto esteja disponível para avaliação. Revisitar e manter essa matriz à medida que o cenário de segurança evolui.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Crie um modelo de ameaça: um modelo de ameaça pode ajudar a identificar e solucionar possíveis ameaças à segurança.
  - NIST: Guide to Data-Centric System Threat Modeling (Guia para modelagem de ameaças de sistemas centrados em dados)

Recursos

Documentos relacionados:

- AWS Security Audit Guidelines (Diretrizes de auditoria de segurança da AWS)
- Boletins de segurança

Vídeos relacionados:

- Security Best Practices the Well-Architected Way

SEC01-BP08 Avaliar e implementar regularmente novos serviços e recursos de segurança

Avalie e implemente serviços e recursos de segurança da AWS e parceiros da AWS que permitem que você desenvolva a postura de segurança da sua workload. O blog de segurança da AWS destaca novos serviços e recursos, guias de implementação e orientações gerais de segurança da

AWS. Quais as novidades da AWS? é uma ótima forma de se manter atualizado com todos os novos recursos, serviços e anúncios da AWS.

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

Orientação de implementação

- Planeje revisões regulares: crie um calendário de atividades de análise que inclua os requisitos de conformidade, avaliar novos recursos e serviços de segurança da AWS e manter-se atualizado sobre as novidades do setor.
- Descubra os serviços e recursos da AWS: descubra os recursos de segurança disponíveis para os serviços que você está usando e analise os novos recursos à medida que são lançados.
  - Blog de segurança da AWS
  - Boletins de segurança da AWS
  - Documentação do serviço da AWS
- Definir processo de integração de serviços da AWS: defina processos para integração de novos serviços da AWS. Inclua como você avalia os novos serviços da AWS em termos de funcionalidade e os requisitos de conformidade para sua workload.
- Teste novos serviços e recursos: teste novos serviços e recursos à medida que são lançados em um ambiente que não seja de produção que replica bem o ambiente de produção.
- Implemente outros mecanismos de defesa: implemente mecanismos automatizados para defender sua workload e explore as opções disponíveis.
  - Como corrigir recursos não compatíveis da AWS pelo Regras do AWS Config

Recursos

Vídeos relacionados:

- Security Best Practices the Well-Architected Way

Gerenciamento de identidade e acesso

Perguntas

- SEC 2 Como você gerencia a autenticação de pessoas e máquinas?
- SEC 3 Como você gerencia permissões para pessoas e máquinas?

SEC 2 Como você gerencia a autenticação de pessoas e máquinas?

Há dois tipos de identidade que você precisa gerenciar para operar workloads 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: seus administradores, desenvolvedores, operadores e usuários finais precisam de uma identidade para acessar seus ambientes e aplicações na AWS. Eles são membros de sua organização ou usuários externos com quem você colabora e que interagem com seus recursos da AWS por meio de um navegador da Web, de uma aplicação cliente ou de ferramentas interativas de linha de comando.

Identidades de máquina: suas aplicações de serviço, ferramentas operacionais e workloads precisam de uma identidade para fazer solicitações a serviços da AWS para ler dados, por exemplo. Essas identidades incluem máquinas em execução em 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.

Práticas recomendadas

- SEC02-BP01 Usar mecanismos de login fortes
- SEC02-BP02 Usar credenciais temporárias
- SEC02-BP03 Armazenar e usar segredos com segurança
- SEC02-BP04 Contar com um provedor de identidades centralizado:
- SEC02-BP05 Fazer a auditoria e a rotação periódica das credenciais
- SEC02-BP06 Utilizar grupos e atributos de usuários

SEC02-BP01 Usar mecanismos de login fortes

Imponha o tamanho mínimo da senha e instrua os usuários a evitar senhas comuns ou reutilizadas. Aplique a Multi-Factor Authentication (MFA – Autenticação multifator) com mecanismos de software ou hardware para fornecer uma camada adicional de verificação. Por exemplo, quando usar o Centro de Identidade do IAM como origem de identidade, defina a configuração de “reconhecimento de contexto” ou “sempre ativo” da MFA e permita que os usuários inscrevam seus próprios dispositivos MFA para acelerar a adoção. Ao usar um Identity Provider (IdP – Provedor de identidade) externo, configure-o para MFA.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Criar uma política do Identify and Access Management (IAM) para aplicar o login de MFA: crie uma política gerenciada pelo cliente do IAM que proíba todas as ações do IAM, exceto aquelas que permitem que um usuário assuma perfis, altere suas próprias credenciais e gerencie seus dispositivos MFA na página My Security Credentials (Minhas credenciais de segurança).
- Habilitar a MFA no provedor de identidades: habilite a MFA no provedor de identidades ou serviço de logon único, como o AWS IAM Identity Center, que você usa.
- Configurar uma política de senhas robusta para seus usuários: configure uma política de senha forte no IAM e nos sistemas de identidade federada para ajudar na proteção contra ataques de força bruta.

- Alternar credenciais regularmenteverifique se os administradores de sua workload alteram senhas e chaves de acesso (se usadas) regularmente.

Recursos

Documentos relacionados:

- Conceitos básicos do AWS Secrets Manager
- Práticas recomendadas do IAM
- Provedores de identidade e federação
- O usuário raiz da conta da AWS
- Conceitos básicos do AWS Secrets Manager
- Credenciais de segurança temporárias
- Soluções para parceiros de segurança: acesso e controle de acesso
- Credenciais de segurança temporárias
- O usuário raiz da conta da AWS

Vídeos relacionados:

- Best Practices for Managing, Retrieving, and Rotating Secrets at Scale (Práticas recomendadas para gerenciar, recuperar e alternar segredos em grande escala)

- Managing user permissions at scale with IAM Identity Center (Gerenciar permissões de usuário em grande escala com o AWS SSO)
- Mastering identity at every layer of the cake

SEC02-BP02 Usar credenciais temporárias

exija que as identidades adquiram credenciais temporárias dinamicamente. Para identidades de força de trabalho, use o AWS IAM Identity Center ou a federação com perfis do AWS Identity and Access Management (IAM) para acessar as Contas da AWS. Para identidades de máquina, como instâncias do Amazon Elastic Compute Cloud(Amazon EC2) ou funções do AWS Lambda, exija o uso de perfis do IAM em vez de usuários do IAM com chaves de acesso de longo prazo.

Para identidades humanas que usam o AWS Management Console, exija que os usuários adquiram credenciais temporárias e façam a federação na AWS. Você pode fazer isso usando o portal do usuário do AWS IAM Identity Center. Para usuários que precisam de acesso à CLI, certifique-se de que eles usem a AWS CLI v2, que oferece suporte para integração direta com o IAM Identity Center. Os usuários podem criar perfis de CLI vinculados a contas e perfis do Centro de Identidade do IAM. A CLI recupera automaticamente as credenciais da AWS do IAM Identity Center e as atualiza em seu nome. Isso elimina a necessidade de copiar e colar credenciais temporárias da AWS no console do IAM Identity Center. Para SDK, os usuários devem contar com o AWS Security Token Service (AWS STS) para assumir perfis e receber credenciais temporárias. Em alguns casos, credenciais temporárias podem não ser práticas. Você deve estar ciente dos riscos de armazenar chaves de acesso. Alterne-as com frequência e exija a autenticação multifator (MFA) como uma condição, quando possível. Use as últimas informações acessadas para determinar quando alternar ou remover as chaves de acesso.

Para casos em que você precisa conceder aos consumidores acesso aos seus recursos da AWS, use os grupos de identidade do Amazon Cognito e atribua a eles um conjunto de credenciais de privilégios temporários e limitados para acessar seus recursos da AWS. As permissões para cada usuário são controladas por meio de perfis do IAM que você cria. Você pode definir regras para escolher a função de cada usuário com base em solicitações no token de ID do usuário. Você pode definir uma função padrão para usuários autenticados. Você também pode definir uma função do IAM separada com permissões limitadas para usuários convidados que não são autenticados.

Para identidades de máquina, você deve confiar em perfis do IAM para conceder acesso à AWS. Para instâncias do Amazon Elastic Compute Cloud(Amazon EC2), você pode usar perfis do Amazon EC2. Você pode anexar um perfil do IAM à sua instância do Amazon EC2 para permitir que suas aplicações em execução no Amazon EC2 usem credenciais de segurança temporárias que a AWS

cria, distribui e alterna automaticamente por meio do Instance Metadata Service (IMDS – Serviço de metadados da instância). A versão mais recente do IMDS ajuda a proteger contra vulnerabilidades que expõem as credenciais temporárias e devem ser implementadas. Para acessar instâncias do Amazon EC2 usando chaves ou senhas, o AWS Systems Manager é uma maneira mais segura de acessar e gerenciar suas instâncias usando um agente pré-instalado sem o segredo armazenado. Além disso, outros serviços da AWS, como o AWS Lambda, permitem que você configure um perfil de serviço do IAM para conceder permissões de serviço a fim de executar ações da AWS usando credenciais temporárias. Em situações em que não é possível usar credenciais temporárias, use ferramentas programáticas, como o AWS Secrets Manager, para automatizar a rotação e o gerenciamento de credenciais.

Fazer a auditoria e a rotação periódica das credenciais: A validação periódica, preferencialmente por meio de uma ferramenta automatizada, é necessária para verificar se os controles corretos são aplicados. Para identidades humanas, você deve exigir que os usuários alterem suas senhas periodicamente e retirem chaves de acesso em favor de credenciais temporárias. Conforme você migra usuários do IAM para identidades centralizadas, é possível gerar um relatório de credenciais para auditar os usuários do IAM. Também recomendamos implementar as configurações de MFA no provedor de identidades. Você pode configurar o Regras do AWS Config para monitorar essas configurações. Para identidades de máquina, você deve confiar em credenciais temporárias usando perfis do IAM. Para situações em que isso não é possível, é necessária a auditoria frequente e a mudança de chaves de acesso.

Armazenar e usar segredos com segurança: para credenciais não relacionadas ao IAM e que não podem usar credenciais temporárias, como logins de banco de dados, use um serviço projetado para lidar com o gerenciamento de segredos, como o Secrets Manager. O Secrets Manager facilita o gerenciamento, a rotação e o armazenamento seguro de segredos criptografados usando serviços com suporte. As chamadas para acessar os segredos são registradas no AWS CloudTrail para fins de auditoria, e as permissões do IAM podem conceder privilégio mínimo a elas.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Implementar políticas de privilégio mínimo: atribua políticas de acesso com privilégio mínimo a grupos e perfis do IAM para refletir a função do usuário ou a função que você definiu.
  - Grant least privilege
- Remover permissões desnecessárias: implemente o privilégio mínimo removendo permissões desnecessárias.

    - Redução do escopo da política ao exibir a atividade do usuário
    - Visualizar acesso à função

- Considerar os limites de permissões: um limite de permissões é um recurso avançado para usar uma política gerenciada que define o número máximo de permissões que uma política baseada em identidade pode conceder a uma entidade do IAM. O limite de permissões de uma entidade permite que ela execute apenas as ações aceitas por suas políticas baseadas em identidade e seus limites de permissões.
    - Laboratório: limites de permissões do IAM que delegam a criação de perfis
- Considerar tags de recursos para permissões: você pode usar tags para controlar o acesso aos recursos da AWS que oferecem suporte à marcação. Você também pode marcar usuários e perfis do IAM para controlar o que eles podem acessar.
    - Laboratório: Controle de acesso baseado em tags do IAM para o EC2
    - AttributeControle de acesso baseado em atributos (ABAC)

Recursos

Documentos relacionados:

- Conceitos básicos do AWS Secrets Manager
- Práticas recomendadas do IAM
- Provedores de identidade e federação
- Soluções para parceiros de segurança: acesso e controle de acesso
- Credenciais de segurança temporárias
- O usuário raiz da conta da AWS

Vídeos relacionados:

- Best Practices for Managing, Retrieving, and Rotating Secrets at Scale (Práticas recomendadas para gerenciar, recuperar e alternar segredos em grande escala)
- Managing user permissions at scale with AWS IAM Identity Center (Gerenciar permissões de usuário em grande escala com o AWS SSO)
- Mastering identity at every layer of the cake

SEC02-BP03 Armazenar e usar segredos com segurança

As identidades de força de trabalho e de máquina 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, como credenciais não relacionadas ao IAM e que não podem usar credenciais temporárias, como logins de banco de dados. Use um serviço projetado para lidar com o gerenciamento de segredos, como o AWS Secrets Manager. O Secrets Manager facilita o gerenciamento, a rotação e o armazenamento seguro de segredos criptografados usando serviços compatíveis. As chamadas para acessar os segredos são registradas no AWS CloudTrail para fins de auditoria, e as permissões do IAM podem conceder acesso de privilégio mínimo a elas.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Use o AWS Secrets Manager: AWS Secrets Manager é um serviço da AWS que facilita o gerenciamento de segredos. Segredos podem ser credenciais de banco de dados, senhas, chaves de API de terceiros e até texto arbitrário.

Recursos

Documentos relacionados:

- Conceitos básicos do AWS Secrets Manager
- Provedores de identidade e federação

Vídeos relacionados:

- Best Practices for Managing, Retrieving, and Rotating Secrets at Scale (Práticas recomendadas para gerenciar, recuperar e alternar segredos em grande escala)

SEC02-BP04 Contar com um provedor de identidades centralizado:

Para identidades da força de trabalho, conte com um provedor de identidade que permita a você gerenciar identidades em um local centralizado. Isso facilita o gerenciamento do acesso em vários aplicativos e serviços, pois você está criando, gerenciando e revogando o acesso de um único local. Por exemplo, se alguém deixar sua organização, você poderá revogar o acesso a todos os serviços e aplicações (incluindo a AWS) de um único local. Esse procedimento reduz a exigência de várias

credenciais e oferece uma oportunidade de integração com processos de recursos humanos (RH) existentes.

Para federação com contas individuais da AWS, você pode usar identidades centralizadas da AWS com um provedor baseado em SAML 2.0 com o AWS Identity and Access Management. Você pode usar qualquer provedor (hospedado por você na AWS, externo à AWS ou fornecido pela AWS Partner, que seja compatível com o protocolo SAML 2.0 . Você pode usar a federação entre sua conta da AWS e o provedor escolhido a fim de conceder acesso a um usuário ou a uma aplicação para chamar operações da API da AWS com uma declaração SAML para obter credenciais de segurança temporárias. Também há suporte para logon único baseado na Web, permitindo que os usuários façam login no AWS Management Console por meio do site de login.

Para federação em várias contas no AWS Organizations, você pode configurar sua origem de identidade no AWS IAM Identity Center (IAM Identity Center)e especificar onde seus usuários e grupos são armazenados. Uma vez configurado, seu provedor de identidade é sua fonte confiável, e as informações podem ser sincronizadaas com o uso do protocolo System for Cross-domain Identity Management (SCIM) v2.0. Em seguida, você pode pesquisar usuários ou grupos e conceder a eles acesso de IAM Identity Center a contas da AWS, aplicações de nuvem ou ambos.

O IAM Identity Center integra-se ao AWS Organizations, o que permite configurar seu provedor de identidade uma vez e, em seguida, conceder acesso a contas novas e existentes gerenciadas na sua organização. O IAM Identity Center fornece um armazenamento padrão, que você pode usar para gerenciar seus usuários e grupos. Se você optar por usar o armazenamento do IAM Identity Center, crie seus usuários e grupos e atribua o nível de acesso deles às suas contas e aplicações da AWS, tendo em mente a prática recomendada do privilégio mínimo. Como alternativa, você pode optar por Conectar-se ao seu provedor de identidade externo usando SAML 2.0 ou Conectar-se ao seu diretório do Microsoft AD usando o AWS Directory Service. Depois de configurado, você pode fazer login no AWS Management Console ou no aplicativo móvel da AWS, autenticando por meio do provedor de identidades central.

Para gerenciar usuários finais ou consumidores de suas cargas de trabalho, como um aplicativo para dispositivos móveis, você pode usar o Amazon Cognito. Ele fornece autenticação, autorização e gerenciamento de usuários para aplicativos Web e para dispositivos móveis. Os usuários podem fazer login diretamente com um nome do usuário e senha ou por meio de terceiros, como Amazon, Apple, Facebook ou Google.

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

Orientação de implementação

- Centralize o acesso administrativo: crie uma entidade de provedor de identidades do Identity and Access Management (IAM) para estabelecer um relacionamento confiável entre o Conta da AWS e o provedor de identidades (IdP). O IAM oferece suporte a IdPs compatíveis com OpenID Connect (OIDC) ou SAML 2.0 (Security Assertion Markup Language 2.0).
  - Provedores de identidade e federação
- Centralize o acesso à aplicação: considere o Amazon Cognito para centralizar o acesso à aplicação. O produto permite que você adicione cadastro/login de usuários e controle de acesso aos seus aplicativos móveis e web de forma rápida e fácil. Amazon Cognito escala para milhões de usuários e oferece suporte ao login com provedores de identidades sociais, como Facebook, Google e Amazon, e provedores de identidade corporativa via SAML 2.0.

- Remova usuários e grupos antigos do IAM: depois de começar a usar um provedor de identidades (IdP), remova usuários e grupos do IAM que não são mais necessários.
  - Encontrar credenciais não utilizadas
  - Excluir um grupo do IAM

Recursos

Documentos relacionados:

- Práticas recomendadas do IAM
- Soluções para parceiros de segurança: acesso e controle de acesso
- Credenciais de segurança temporárias
- O usuário raiz da conta da AWS

Vídeos relacionados:

- Best Practices for Managing, Retrieving, and Rotating Secrets at Scale (Práticas recomendadas para gerenciar, recuperar e alternar segredos em grande escala)
- Managing user permissions at scale with AWS IAM Identity Center (Gerenciar permissões de usuário em grande escala com o AWS SSO)
- Mastering identity at every layer of the cake

SEC02-BP05 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, autenticação multifator (MFA), sejam aplicados, alternados regularmente e que tenham o nível de acesso apropriado. A validação periódica, preferencialmente por meio de uma ferramenta automatizada, é necessária para verificar se os controles corretos são aplicados. Para identidades humanas, você deve exigir que os usuários alterem suas senhas periodicamente e retirem chaves de acesso em favor de credenciais temporárias. Conforme você migra usuários do AWS Identity and Access Management (IAM) para identidades centralizadas, é possível gerar um relatório de credenciais para auditar os usuários do IAM. Também recomendamos que implementar as configurações de MFA no provedor de identidades. Você pode configurar o Regras do AWS Config para monitorar essas configurações. Para identidades de máquina, você deve confiar em credenciais temporárias usando perfis do IAM. Para situações em que isso não é possível, é necessária a auditoria frequente e a mudança de chaves de acesso.

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

Orientação de implementação

- Faça auditoria de credenciais regularmente: use relatórios de credenciais e o Identity and Access Management (IAM) Access Analyzer para auditar credenciais e permissões do IAM.
  - IAM Access Analyzer
  - Obtenção do relatório de credenciais
  - Laboratório: Limpeza automatizada de usuários do IAM
- Use os níveis de acesso para revisar as permissões do IAM: para melhorar a segurança da sua Conta da AWS, revise e monitore regularmente cada uma das políticas do IAM. Certifique-se de que suas políticas concedam o privilégio mínimo para executar apenas as ações necessárias.
  - Usar níveis de acesso para revisar permissões do IAM
- Considere automatizar a criação e as atualizações de recursos do IAM: o AWS CloudFormation pode ser usado para automatizar a implantação de recursos do IAM, incluindo perfis e políticas, para reduzir erros humanos, pois os modelos podem ser verificados e ter controle de versão.
  - Laboratório: Implantação automatizada de grupos e perfis do IAM

Recursos

Documentos relacionados:

- Conceitos básicos do AWS Secrets Manager
- Práticas recomendadas do IAM
- Provedores de identidade e federação
- Soluções para parceiros de segurança: acesso e controle de acesso
- Credenciais de segurança temporárias

Vídeos relacionados:

- Best Practices for Managing, Retrieving, and Rotating Secrets at Scale (Práticas recomendadas para gerenciar, recuperar e alternar segredos em grande escala)
- Managing user permissions at scale with AWS IAM Identity Center (Gerenciar permissões de usuário em grande escala com o AWS SSO)
- Mastering identity at every layer of the cake

SEC02-BP06 Utilizar grupos e atributos de usuários

À medida que o número de usuários gerenciados cresce, você precisará determinar maneiras de organizá-los para que você possa gerenciá-los em grande escala. 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 para controlar o acesso em vez de usuários individuais. Isso permite que você gerencie o acesso centralmente, alterando a associação ao grupo ou os atributos de um usuário uma vez com um conjunto de permissões, em vez de atualizar várias políticas individuais quando as necessidades de acesso de um usuário mudarem. Você pode usar o AWS IAM Identity Center (IAM Identity Center) para gerenciar grupos e atributos de usuários. O IAM Identity Center oferece suporte aos atributos mais usados, quer eles sejam inseridos manualmente durante a criação do usuário ou provisionados automaticamente usando um mecanismo de sincronização, como definido na especificação System for Cross-Domain Identity Management (SCIM).

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.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Se estiver usando o AWS IAM Identity Center (IAM Identity Center), configure grupos: o IAM Identity Center permite configurar grupos de usuários e atribuir aos grupos o nível desejado de permissão.
  - AWS Single Sign-On: gerenciar identidades
- Saiba mais sobre o controle de acesso por atributo (ABAC): o ABAC é uma estratégia de autorização que define permissões com base em atributos.
  - O que é ABAC para a AWS?
  - Laboratório: Controle de acesso baseado em tags do IAM para o EC2

Recursos

Documentos relacionados:

- Conceitos básicos do AWS Secrets Manager
- Práticas recomendadas do IAM
- Provedores de identidade e federação
- O usuário raiz da conta da AWS

Vídeos relacionados:

- Best Practices for Managing, Retrieving, and Rotating Secrets at Scale (Práticas recomendadas para gerenciar, recuperar e alternar segredos em grande escala)
- Managing user permissions at scale with AWS IAM Identity Center (Gerenciar permissões de usuário em grande escala com o AWS SSO)
- Mastering identity at every layer of the cake

Exemplos relacionados:

- Laboratório: Controle de acesso baseado em tags do IAM para o EC2

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 workload. As permissões controlam quem pode acessar o quê e em quais condições.

Práticas recomendadas

- SEC03-BP01 Definir requisitos de acesso
- SEC03-BP02 Conceder acesso com privilégio mínimo
- SEC03-BP03 Estabelecer processo de acesso de emergência
- SEC03-BP04 Reduzir as permissões continuamente
- SEC03-BP05 Definir barreiras de proteção de permissões para sua organização
- SEC03-BP06 Gerenciar o acesso com base no ciclo de vida
- SEC03-BP07 Analisar o acesso público e entre contas
- SEC03-BP08 Compartilhar recursos com segurança

SEC03-BP01 Definir requisitos de acesso

Cada componente ou recurso de sua workload 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, escolher o tipo de identidade apropriado e o método de autenticação e autorização.

Antipadrões comuns:

- Codificação rígida ou armazenamento de segredos em sua aplicação.
- Conceder permissões personalizadas a cada usuário.
- Uso de credenciais de longa duração.

Nível de risco exposto se essa prática recomendada não for estabelecida: alto

Orientação para implementação

Cada componente ou recurso de sua workload 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, escolher o tipo de identidade apropriado e o método de autenticação e autorização.

O acesso regular a Contas da AWS na organização deve ser fornecido usando acesso federado ou um provedor de identidade centralizado. Você também deve centralizar o gerenciamento de identidades e garantir que haja uma prática estabelecida para integrar o acesso à AWS ao ciclo de vida de acesso dos funcionários. Por exemplo, quando um funcionário muda para um cargo com um nível de acesso diferente, sua associação ao grupo também deve mudar para refletir os novos requisitos de acesso.

Ao definir os requisitos de acesso para identidades não humanas, determine quais aplicações e componentes precisam de acesso e como as permissões são concedidas. O uso de perfis do IAM criados com o modelo de acesso de privilégio mínimo é uma abordagem recomendada. As políticas gerenciadas pela AWS fornecem políticas predefinidas do IAM que abordam a maioria dos casos de uso comuns.

Os serviços da AWS, como o AWS Secrets Manager e o AWS Systems Manager Parameter Store, podem ajudar a desacoplar segredos da aplicação ou workload com segurança em casos em que não é possível usar perfis do IAM. No Secrets Manager, você pode estabelecer uma alternância automática de suas credenciais. É possível usar o Systems Manager para referenciar parâmetros em seus scripts, comandos, documentos do SSM, configurações e fluxos de trabalho de automação, usando o nome exclusivo que você especificou ao criar o parâmetro.

Você pode usar o AWS Identity and Access Management Roles Anywhere para obter credenciais de segurança temporárias no IAM para workloads executadas fora da AWS. As workloads podem usar as mesmas políticas do IAM e perfis do IAM que você usa com as aplicações da AWS para acessar os recursos da AWS.

Quando possível, prefira credenciais temporárias de curta duração em vez de credenciais estáticas de longa duração. Para cenários em que você precisa de usuários da IAM com acesso programático e credenciais de longa duração, use as últimas informações usadas da chave de acesso para alternar e remover chaves de acesso.

Recursos

Documentos relacionados:

- Controle de acesso por atributo (ABAC)
- AWS IAM Identity Center
- IAM Roles Anywhere
- AWS Managed policies for IAM Identity Center (Políticas gerenciadas pela AWS para o IAM Identity Center)
- AWS IAM policy conditions (Condições de políticas do AWS IAM)
- IAM use cases (Casos de uso do IAM)
- Remova credenciais desnecessárias
- Trabalhando com políticas
- How to control access to AWS resources based on Conta da AWS, OU, or organization (Como controlar o acesso aos recursos da AWS baseados em Conta da AWS, UO ou organização)
- Identify, arrange, and manage secrets easily using enhanced search in AWS Secrets Manager (Identificar, organizar e gerenciar segredos facilmente usando a pesquisa avançada no AWS Secrets Manager)

Vídeos relacionados:

- Become an IAM Policy Master in 60 Minutes or Less (Torne-se um mestre em políticas do IAM em 60 minutos ou menos)
- Separation of Duties, Least Privilege, Delegation, and CI/CD (Separação de tarefas, privilégio mínimo, delegação e CI/CD)
- Streamlining identity and access management for innovation (Simplificação do gerenciamento de identidade e acesso para inovação)

SEC03-BP02 Conceder acesso com privilégio mínimo

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.

Antipadrões comuns:

- Usar como padrão a concessão de permissões de administrador aos usuários.
- Usar a conta raiz para atividades diárias.

Nível de risco exposto se essa prática recomendada não for estabelecida: alto

Orientação para implementação

Estabelecer um princípio de privilégio mínimo garante que as identidades só tenham permissão para executar o conjunto mínimo de funções necessárias para realizar uma tarefa específica, enquanto equilibram usabilidade e eficiência. Operar com esse princípio limita o acesso não intencional e ajuda a garantir que você possa auditar quem tem acesso a quais recursos. Na AWS, as identidades não têm permissões por padrão, exceto para o usuário raiz. As credenciais do usuário raiz devem ser estritamente controladas e só podem ser usadas para algumas tarefas específicas.

Você usa políticas para conceder explicitamente permissões anexadas ao IAM ou a entidades de recursos, como um perfil do IAM usado por máquinas ou identidades federadas, ou recursos (por exemplo, buckets do S3). Ao criar e associar uma política, você pode especificar as ações de serviço, os recursos e as condições que devem ser verdadeiros para que a AWS permita o acesso. A AWS oferece suporte a uma variedade de condições para ajudar você a reduzir o acesso. Por exemplo, usando a chave de condição `PrincipalOrgID` ,o identificador do AWS Organizations é verificado para que o acesso possa ser concedido dentro do AWS Organization.

Você também pode controlar as solicitações feitas pelos serviços da AWS em seu nome, como o AWS CloudFormation criando uma função do AWS Lambda, usando a chave de condição `CalledVia` . Você deve colocar em camadas diferentes tipos de política para limitar efetivamente as permissões gerais em uma conta. Por exemplo, é possível permitir que suas equipes de aplicação criem suas próprias políticas do IAM, mas usar um limite de permissões para definir o máximo de permissões que elas podem conceder.

Há vários recursos da AWS para ajudar a escalar o gerenciamento de permissões e aderir ao princípio do privilégio mínimo. O controle de acesso baseado em atributos permite limitar as permissões com base na tag de um recurso, visando tomar decisões de autorização de acordo com as tags aplicadas ao recurso e a chamada de uma entidade principal do IAM. Isso permite combinar sua política de permissões e marcação para obter um acesso refinado a recursos sem precisar de muitas políticas personalizadas.

Outra maneira de acelerar a criação de uma política de privilégio mínimo é basear sua política nas permissões do CloudTrail depois da execução de uma atividade. O IAM Access Analyzer pode gerar automaticamente uma política do IAM baseada na atividade. Também é possível usar o IAM

Access Advisor no nível da organização ou da conta individual para monitorar as últimas informações acessadas de uma política específica.

Estabeleça uma frequência para revisar esses detalhes e remover permissões desnecessárias. Você deve estabelecer uma barreira de proteção de permissões na organização da AWS para controlar o máximo de permissões na conta de qualquer membro. Serviços como o AWS Control Tower têm controles preventivos, gerenciados e prescritivos e permitem definir seus próprios controles.

Recursos

Documentos relacionados:

- Permissions boundaries for IAM entities (Limites de permissões para entidades do IAM)
- Techniques for writing least privilege IAM policies (Técnicas para escrever políticas do IAM de privilégio mínimo)
- IAM Access Analyzer makes it easier to implement least privilege permissions by generating IAM policies based on access activity (IAM Access Analyzer facilita a implementação de permissões de privilégio mínimo gerando políticas do IAM baseadas na atividade de acesso)
- Refining Permissions using last accessed information (Refinar permissões usando as últimas informações acessadas)
- IAM policy types and when to use them (Tipos de política do IAM e quando usá-las)
- Testing IAM policies with the IAM policy simulator (Testar políticas do IAM com o simulador de política do IAM)
- Guardrails in AWS Control Tower (Barreiras de proteção no AWS Control Tower)
- Zero Trust architectures: An AWS perspective (Arquiteturas de confiança zero: uma perspectiva da AWS)
- How to implement the principle of least privilege with CloudFormation StackSets (Como implementar o princípio de privilégio mínimo com o CloudFormation StackSets)

Vídeos relacionados:

- Next-generation permissions management (Gerenciamento de permissões de última geração)
- Zero Trust: An AWS perspective (Confiança zero: uma perspectiva da AWS)
- How can I use permissions boundaries to limit IAM users and roles to prevent privilege escalation? (Como posso usar limites de permissões para limitar usuários e perfis do IAM para evitar a escalação de privilégios?)

Exemplos relacionados:

- Lab: IAM permissions boundaries delegating role creation (Laboratório: limites de permissões do IAM que delegam a criação de perfis)

SEC03-BP03 Estabelecer processo de acesso de emergência

Um processo que permite o acesso de emergência à sua workload 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. Esse processo pode incluir uma combinação de recursos diferentes, por exemplo, um perfil de emergência entre contas da AWS para acesso ou um processo específico para os administradores seguirem para validar e aprovar uma solicitação de emergência.

Antipadrões comuns:

- Não ter um processo de emergência vigente para se recuperar de uma interrupção com sua configuração de identidade existente.
- Conceder permissões elevadas de longa duração para fins de recuperação ou resolução de problemas.

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

Orientação para implementação

O estabelecimento de um acesso de emergência pode assumir diversos formatos para os quais você deve estar preparado. O primeiro é uma falha de seu provedor de identidades primário. Nesse caso, você deve utilizar um segundo método de acesso com as permissões necessárias para a recuperação. Esse método pode ser um provedor de identidade de backup ou um usuário do IAM. Esse segundo método deve ser estritamente controlado, monitorado e notificado caso seja usado. A identidade de acesso de emergência deve ser originada de uma conta específica para esse fim e só deve ter permissões para assumir um perfil especificamente projetado para recuperação.

Você também deverá se preparar para o acesso de emergência quando o acesso administrativo elevado temporário for necessário. Um cenário comum é limitar as permissões mutantes a um processo automatizado usado para implantar modificações. Se esse processo apresentar um problema, os usuários podem precisar solicitar permissões elevadas para restaurar a funcionalidade. Nesse caso, estabeleça um processo em que os usuários possam solicitar acesso elevado e os administradores possam validá-lo e aprová-lo. Os planos de implementação detalhando as

orientações de práticas recomendadas para funções com acesso pré-provisionado e preparação para emergências, break-glass, são fornecidos como parte do SEC10-BP05 Acesso pré-provisionado.

Recursos

Documentos relacionados:

- Monitor and Notify on AWS (Monitoramento e notificação na AWS)
- Managing temporary elevated access (Gerenciamento do acesso elevado temporário)

Vídeo relacionado:

- Become an IAM Policy Master in 60 Minutes or Less (Torne-se um mestre em políticas do IAM em 60 minutos ou menos)

SEC03-BP04 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.

Às vezes, quando equipes e projetos estão apenas começando, você pode optar por conceder amplo acesso (em um ambiente de desenvolvimento ou teste) para inspirar inovação e agilidade. Recomendamos avaliar o acesso continuamente e, particularmente em um ambiente de produção, restrinja o acesso apenas às permissões necessárias e obtenha privilégio mínimo. A AWS fornece recursos de análise de acesso para ajudar a identificar o acesso não utilizado. Para ajudar a identificar usuários, funções, permissões e credenciais não utilizados, a AWS analisa a atividade de acesso e fornece informações sobre a chave de acesso e a função usadas mais recentemente. Você pode usar o timestamp de último acesso to identificar usuários e funções não utilizadose removê-los. Além disso, você pode revisar as informações de último acesso a serviços e ações para identificar e restringir permissões para usuários e funções específicos. Por exemplo, você pode usar as informações acessadas mais recentemente para identificar as ações específicas do Amazon Simple Storage Service(Amazon S3) exigidas pela função da aplicação e restringir o acesso apenas a essas ações. Esses recursos estão disponíveis no AWS Management Console e de maneira programática para permitir que você os incorpore aos fluxos de trabalho de infraestrutura e ferramentas automatizadas.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Configure o AWS Identify and Access Management (IAM) Access Analyzer: o AWS IAM Access Analyzer ajuda você a identificar os recursos na organização e nas contas, como buckets do Amazon Simple Storage Service (Amazon S3) ou funções do IAM, que são compartilhados com uma entidade externa.
  - AWS IAM Access Analyzer

Recursos

Documentos relacionados:

- AttributeControle de acesso baseado em atributos (ABAC)
- Grant least privilege
- Remova credenciais desnecessárias
- Trabalhando com políticas

Vídeos relacionados:

- Become an IAM Policy Master in 60 Minutes or Less (Torne-se um mestre em políticas do IAM em 60 minutos ou menos)
- Separation of Duties, Least Privilege, Delegation, and CI/CD (Separação de tarefas, privilégio mínimo, delegação e CI/CD)

SEC03-BP05 Definir barreiras de proteção de permissões para sua organização

Estabeleça controles comuns que restrinjam o acesso a todas as identidades na organização. Por exemplo, é possível restringir o acesso a Regiões da AWS específicas ou impedir que os operadores excluam recursos comuns, como um perfil do IAM usado pela equipe de segurança central.

Antipadrões comuns:

- Execução de workloads em sua conta de administrador organizacional.
- Execução de workloads de produção e não produção na mesma conta.

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

Orientação para implementação

Com a expansão e o gerenciamento de workloads adicionais na AWS, você deve separá-las usando contas e gerenciá-las usando o AWS Organizations. Recomendamos que você estabeleça barreiras de proteção de permissões comuns que restrinjam o acesso a todas as identidades na sua organização. Por exemplo, você pode restringir o acesso a Regiões da AWS específicas ou impedir que a equipe exclua recursos comuns, como um perfil do IAM usado pela equipe de segurança central.

Você pode começar implementando exemplos de políticas de controle de serviço, como impedir que os usuários desabilitem os principais serviços. As SCPs usam a linguagem de políticas do IAM e permitem que você estabeleça controles aos quais todas as entidades principais (usuários e perfis) do IAM aderem. Você pode restringir o acesso a ações de serviço, recursos específicos e com base em condições específicas para atender às necessidades de controle de acesso de sua organização. Se necessário, você pode definir exceções para suas barreiras de proteção. Por exemplo, você pode restringir ações de serviço para todas as entidades do IAM na conta, exceto para um perfil de administrador específico.

Recomendamos evitar a execução de workloads em sua conta de gerenciamento. A conta de gerenciamento deve ser usada para gerir e implantar barreiras de proteção de segurança que afetarão as contas-membro. Alguns serviços da AWS permitem o uso de uma conta de administrador delegada. Quando disponível, você deve usar essa conta delegada em vez da conta de gerenciamento. Você deve limitar estritamente o acesso à conta de administrador organizacional.

O uso de uma estratégia de várias contas permite ter maior flexibilidade na aplicação de barreiras de proteção às suas workloads. O AWS Security Reference Architecture dá orientações prescritivas sobre como projetar a estrutura da conta. Os serviços da AWS, como o AWS Control Tower, fornece recursos para gerenciar centralmente os controles de prevenção e detecção em sua organização. Defina um objetivo claro para cada conta ou UO em sua organização e limite os controles de acordo com esse objetivo.

Recursos

Documentos relacionados:

- AWS Organizations
- Service control policies (SCPs) (Políticas de controle de serviços (SCPs))
- Get more out of service control policies in a multi-account environment (Aproveite ao máximo as políticas de controle de serviços em um ambiente de várias contas)

- AWS Security Reference Architecture (AWS SRA)

Vídeos relacionados:

- Enforce Preventive Guardrails using Service Control Policies (Aplique barreiras de proteção preventivas usando políticas de controle de serviços)
- Building governance at scale with AWS Control Tower (Criação de governança em escala com o AWS Control Tower)
- AWS Identity and Access Management deep dive (Análise aprofundada do AWS Identity and Access Management)

SEC03-BP06 Gerenciar o acesso com base no ciclo de vida

Integre controles de acesso ao ciclo de vida do operador e da aplicação e ao seu provedor de federação centralizado. Por exemplo, remova o acesso do usuário que sair da organização ou mudar de funções.

À medida que você gerencia cargas de trabalho usando contas separadas, haverá casos em que você precisará compartilhar recursos entre essas contas. Recomendamos que você compartilhe recursos usando o AWS Resource Access Manager (AWS RAM). Esse serviço permite que você compartilhe, com facilidade e segurança, os recursos da AWS dentro da AWS Organizations e das unidades organizacionais. Usando o AWS RAM, o acesso a recursos compartilhados é concedido ou revogado automaticamente à medida que as contas são movidas para dentro e para fora da organização ou da unidade organizacional com a qual são compartilhadas. Isso ajuda a garantir que os recursos sejam compartilhados apenas com as contas que você determinar.

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

Orientação de implementação

Ciclo de vida de acesso de usuário: implemente uma política de ciclo de vida de acesso para novos usuários, alterações de função de trabalho e usuários que saem, para que apenas os usuários atuais tenham acesso.

Recursos

Documentos relacionados:

- AttributeControle de acesso baseado em atributos (ABAC)

- Grant least privilege
- IAM Access Analyzer
- Remova credenciais desnecessárias
- Trabalhando com políticas

Vídeos relacionados:

- Become an IAM Policy Master in 60 Minutes or Less (Torne-se um mestre em políticas do IAM em 60 minutos ou menos)
- Separation of Duties, Least Privilege, Delegation, and CI/CD (Separação de tarefas, privilégio mínimo, delegação e CI/CD)

SEC03-BP07 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 somente aos recursos que exigem esse tipo de acesso.

Antipadrões comuns:

- Não seguir um processo para gerir o acesso público e entre contas aos recursos.

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

Orientação para implementação

Na AWS, você pode conceder acesso a recursos em outra conta. Você concede acesso direto entre contas usando políticas anexadas a recursos (por exemplo, políticas de bucket do Amazon Simple Storage Service (Amazon S3)) ou permitindo que uma identidade assuma um perfil do IAM em outra conta. Ao usar políticas de recursos, verifique o acesso concedido a identidades em sua organização e se você tem a intenção de tornar os recursos públicos. Defina um processo para aprovar todos os recursos que devem ser acessíveis publicamente.

O IAM Access Analyzer usa segurança comprovada para identificar todos os caminhos de acesso a um recurso de fora de sua conta. Ele revisa as políticas de recursos continuamente e relata descobertas de acesso público e entre contas para facilitar a análise de acesso potencialmente amplo. Considere a configuração do IAM Access Analyzer com o AWS Organizations para verificar se você tem visibilidade em todas as suas contas. O IAM Access Analyzer também permite visualizar

as descobertas do Access Analyzerantes de implantar as permissões do recurso. Isso permite validar que as alterações de política concedam apenas o acesso público e entre contas pretendido aos seus recursos. Ao projetar o acesso de várias contas, é possível usar políticas de confiança para controlar em quais casos um perfil pode ser assumido. Por exemplo, você pode limitar que um perfil seja assumido por determinado intervalo de IPs de origem.

Você também pode usar o AWS Config para relatar e corrigir recursos com uma configuração acidental de acesso público por meio de verificações de políticas do AWS Config. Serviços como o AWS Control Tower e o AWS Security Hub simplificam as barreiras de proteção e as verificações de implantação em uma AWS Organizations para identificar e corrigir recursos publicamente expostos. Por exemplo, o AWS Control Tower tem uma barreira de proteção gerenciada que pode detectar se algum snapshot do Amazon EBS pode ser restaurado por todas as contas da AWS.

Recursos

Documentos relacionados:

- Using AWS Identity and Access Management Access Analyzer (Uso do AWS Identity and Access Management Access Analyzer)
- Guardrails in AWS Control Tower (Barreiras de proteção no AWS Control Tower)
- AWS Foundational Security Best Practices standard (Norma de práticas de segurança básicas da AWS)
- AWS Config Managed Rules (Regras gerenciadas do AWS Config)
- AWS Trusted Advisor check reference (Referência de verificação do AWS Trusted Advisor)

Vídeos relacionados:

- Best Practices for securing your multi-account environment (Práticas recomendadas para proteger seu ambiente de várias contas)
- Dive Deep into IAM Access Analyzer (Análise aprofundada do IAM Access Analyzer)

SEC03-BP08 Compartilhar recursos com segurança

Controle o consumo de recursos compartilhados entre contas ou no AWS Organizations. Monitore recursos compartilhados e revise o acesso a recursos compartilhados.

Antipadrões comuns:

- Uso da política de confiança padrão do IAM ao conceder acesso entre contas de terceiros.

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

Orientação para implementação

Como você gerencia as workloads usando várias contas da AWS, pode ser necessário compartilhar recursos entre contas. Isso será frequentemente um compartilhamento entre contas em uma AWS Organizations. Vários serviços da AWS, como o AWS Security Hub, o Amazon GuardDutye o AWS Backup têm recursos entre contas integrados à Organizations. Você pode usar o AWS Resource Access Manager para compartilhar outros recursos comuns, como sub-redes de VPC ou anexos do gateway de trânsito, o AWS Network Firewallou pipelines do Amazon SageMaker Runtime. Se você quiser garantir que sua conta compartilhe recursos somente com sua Organizations, recomendamos o uso de Service control policies (SCPs) (Políticas de controle de serviços (SCPs)) para impedir o acesso a entidades principais externas.

Ao compartilhar recursos, você deve implantar medidas para se proteger contra acessos indesejados. Recomendamos combinar controles baseados em identidade e controles de rede para criar um perímetro de dados para sua organização. Esses controles devem impor limites estritos sobre quais recursos podem ser compartilhados e impedir o compartilhamento ou a exposição de recursos que não devem ser permitidos. Por exemplo, como parte de seu perímetro de dados, você pode usar políticas de endpoint da VPC e a condição `aws:PrincipalOrgId` para garantir que as identidades acessem os buckets do Amazon S3 pertencentes à sua organização.

Em alguns casos, você pode compartilhar recursos fora de sua Organizations ou conceder a terceiros acesso à sua conta. Por exemplo, um parceiro pode fornecer uma solução de monitoramento que precise acessar recursos em sua conta. Nesses casos, você deve criar um perfil entre contas do IAM somente com os privilégios necessários para a parte externa. Você deve também criar uma política de confiança usando a condição de ID externo. Ao usar um ID externo, você deve gerar um ID exclusivo para cada parte externa. O ID exclusivo não deve ser fornecido nem controlado por essa parte. Se ela não precisar mais de acesso ao seu ambiente, remova o perfil. Você também deve evitar o fornecimento de credenciais do IAM de longa duração para terceiros em todos os casos. Esteja ciente de outros serviços da AWS que sejam compatíveis nativamente com o compartilhamento. Por exemplo, o AWS Well-Architected Tool permite compartilhar uma workload com outras contas da AWS.

Ao usar um serviço como o Amazon S3, é recomendável desabilitar as ACLs para seu bucket do Amazon S3 e usar políticas do IAM para definir o controle de acesso. Para restringir o acesso a uma

origem do Amazon S3 pelo Amazon CloudFront, migre da identidade do acesso de origem (OAI) para um controle de acesso de origem (OAC), que é compatível com recursos adicionais, incluindo a criptografia do lado do servidor com o AWS KMS.

Recursos

Documentos relacionados:

- Bucket owner granting cross-account permission to objects it does not own (Proprietário do bucket concede permissão entre contas a objetos que não possui)
- How to use Trust Policies with IAM (Como usar políticas de confiança com o IAM)
- Building Data Perimeter on AWS (Como criar um perímetro de dados na AWS)
- How to use an external ID when granting a third party access to your AWS resources (Como usar um ID externo ao conceder acesso aos seus recursos da AWS para terceiros)

Vídeos relacionados:

- Granular Access with AWS Resource Access Manager (Acesso granular com o AWS Resource Access Manager)
- Securing your data perimeter with VPC endpoints (Como proteger seu perímetro de dados com endpoints da VPC)
- Establishing a data perimeter on AWS (Como estabelecer um perímetro de dados na AWS)

Detecção

Pergunta

- SEC 4 Como você detecta e investiga eventos de segurança?

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.

Práticas recomendadas

- SEC04-BP01 Configurar registro em log de serviço e aplicação
- SEC04-BP02 Analisar logs, descobertas e métricas de forma centralizada

- SEC04-BP03 Automatizar a resposta a eventos
- SEC04-BP04 Implementar eventos de segurança acionáveis

SEC04-BP01 Configurar registro em log de serviço e aplicação

Configure o registro em log em toda a workload, incluindo logs de aplicações, 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 sua organização.

Uma prática básica é estabelecer um conjunto de mecanismos de detecção no nível da conta. Esse conjunto básico de mecanismos deve registrar e detectar uma grande variedade de ações em todos os recursos da conta. Eles permitem criar uma função de detecção abrangente com opções que incluem correção automatizada e integrações de parceiros para funcionalidade adicional.

Na AWS, os serviços que podem implementar esse conjunto base incluem:

- AWS CloudTrail fornece histórico de eventos da atividade de sua conta da AWS, incluindo ações realizadas por meio do AWS Management Console, de AWS SDKs, de ferramentas de linha de comando e de outros serviços da AWS.
- AWS Config monitora e registra as configurações de recursos da AWS e permite automatizar as tarefas de avaliação e correção em relação às configurações desejadas.
- Amazon GuardDuty é um serviço de detecção de ameaças que monitora continuamente atividades maliciosas e comportamentos não autorizados para proteger contas e cargas de trabalho da AWS.
- AWS Security Hub fornece um único local que agrega, organiza e prioriza alertas de segurança ou descobertas de vários serviços da AWS e produtos opcionais de terceiros para oferecer uma visão abrangente dos alertas de segurança e do status de conformidade.

Com base nos alicerces no nível da conta, muitos serviços essenciais da AWS, como o Amazon Virtual Private Cloud Console (Amazon VPC), fornecem recursos de registro em log em nível de serviço. Logs de fluxo da Amazon VPC permitem capturar informações sobre o tráfego de IP de entrada e saída das interfaces de rede que podem ser valiosas para o histórico de conectividade, além de acionar ações automatizadas com base em comportamentos anômalos.

Para instâncias do Amazon Elastic Compute Cloud(Amazon EC2) e registro em log baseado em aplicações que não são originadas de serviços da AWS, os logs podem ser armazenados e analisados com o Amazon CloudWatch Logs. Uma agente coleta os logs no sistema operacional

e nos aplicativos em execução e os armazena automaticamente. Assim que os logs estiverem disponíveis no CloudWatch Logs, você poderá processá-los em tempo realou se aprofundar em análises usando o CloudWatch Logs Insights.

Igualmente importante para coletar e agregar logs é a capacidade de extrair informações relevantes dos grandes volumes de dados de log e eventos gerados por arquiteturas modernas e complexas. Consulte a guia Monitoramento do whitepaper sobre o pilar de confiabilidade para obter mais detalhes. Os logs podem conter dados considerados confidenciais. Quando os dados do aplicativo são erroneamente encontrados em arquivos de log que o agente do CloudWatch Logs está capturando ou quando o registro em log entre regiões está configurado para agregação de logs e há considerações legislativas sobre o envio de determinados tipos de informações além de fronteiras.

Uma abordagem é usar funções do AWS Lambda, acionadas em eventos quando os logs são entregues, para filtrar e redigir dados de log antes de encaminhá-los para um local de registro centralizado de logs, como um bucket do Amazon Simple Storage Service (Amazon S3). Os logs não editados podem ser mantidos em um bucket local por um tempo razoável (conforme determinado pela legislação e a equipe jurídica), quando uma regra de ciclo de vida do Amazon S3 pode excluí-los automaticamente. Os logs podem ser protegidos ainda mais no Amazon S3 usando o bloqueio de objetos do Amazon S3, no qual é possível armazenar objetos usando um modelo de gravação única e leitura múltipla (WORM).

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Habilitar o registro em log de serviços da AWS: habilite o registro em log de serviços da AWS para atender aos seus requisitos. Os recursos de registro em log incluem o seguinte: logs de fluxo do Amazon VPC, logs do Elastic Load Balancing (ELB), logs de bucket do Amazon S3, logs de acesso do CloudFront, logs de consulta do Amazon Route 53 e logs do Amazon Relational Database Service (Amazon RDS).
  - AWS Answers: capacidade nativa de log de segurança da AWS
- Avalie e habilite o registro em log de sistemas operacionais e logs específicos de aplicativos para detectar comportamentos suspeitos.
  - Conceitos básicos do CloudWatch Logs
  - Ferramentas do desenvolvedor e análise de log
- Aplicar os controles apropriados aos logs: os logs podem conter informações confidenciais e somente usuários autorizados devem ter acesso. Considere restringir as permissões aos grupos de logs dos buckets do Amazon S3 e do CloudWatch Logs.

  - Autenticação e controle de acesso para o Amazon CloudWatch
  - Identity and Access Management no Amazon S3.
- Configurar Amazon GuardDuty: o GuardDuty é um serviço de detecção de ameaças que monitora continuamente atividades maliciosas e comportamentos não autorizados para proteger contas e workloads das Contas da AWS. Habilite o GuardDuty e configure alertas automatizados para enviar e-mails usando o laboratório.
- Configurar trilha personalizada no CloudTrail: a configuração de uma trilha permite armazenar logs por mais tempo que o período padrão e analisá-los posteriormente.
- Habilitar AWS Config: o AWS Config oferece uma visualização detalhada da configuração dos recursos da AWS em uma Conta da AWS. Isso inclui como os recursos se relacionam entre si e como foram configurados anteriormente, permitindo que você veja como as configurações e os relacionamentos mudam ao longo do tempo.
- Habilitar AWS Security Hub: o Security Hub fornece uma visão abrangente do seu estado de segurança na AWS e ajuda a verificar sua conformidade com os padrões e práticas recomendadas do setor de segurança. O Security Hub coleta dados de segurança de todas as Contas da AWS, serviços e produtos de parceiros de terceiros suportados e ajuda você a analisar suas tendências de segurança e identificar os problemas de segurança de maior prioridade.

Recursos

Documentos relacionados:

- Amazon CloudWatch
- Amazon EventBridge
- Conceitos básicos: Amazon CloudWatch Logs
- Soluções de segurança parceiros: registro em log e monitoramento

Vídeos relacionados:

- Centrally Monitoring Resource Configuration and Compliance (Monitoramento centralizado de configuração e conformidade de recursos)
- Remediating Amazon GuardDuty and AWS Security Hub Findings (Correção do Amazon GuardDuty e descobertas do AWS Security Hub)
- Threat management in the cloud: Amazon GuardDuty and AWS Security Hub (Gerenciamento de ameaças na nuvem: Amazon GuardDuty e AWS Security Hub)

Exemplos relacionados:

- Laboratório: Implantação automatizada de controles de detecção

SEC04-BP02 Analisar logs, descobertas e métricas de forma centralizada

as equipes de operações de segurança confiam na coleta de logs e no uso de ferramentas de pesquisa para descobrir possíveis eventos de interesse, que podem indicar atividade não autorizada ou alteração não intencional. No entanto, a simples análise de dados coletados e o processamento manual de informações são insuficientes para acompanhar o volume de informações provenientes de arquiteturas complexas. Somente a análise e os relatórios não facilitam a atribuição dos recursos certos para trabalhar um evento em tempo hábil.

Uma prática recomendada para montar uma equipe madura de operações de segurança é integrar profundamente o fluxo de eventos e descobertas de segurança em um sistema de notificação e fluxo de trabalho, como um sistema de emissão de tíquetes, um sistema de erros ou problemas, ou outro sistema de gerenciamento de informações e eventos de segurança (SIEM). Isso remove o fluxo de trabalho de e-mails e relatórios estáticos, o que permite rotear, escalar e gerenciar eventos ou descobertas. Muitas organizações também estão integrando alertas de segurança em suas plataformas de bate-papo ou colaboração e de produtividade do desenvolvedor. Para organizações que estão iniciando com automações, um sistema de emissão de tíquetes orientado por APIs e de baixa latência oferece flexibilidade considerável para o planejamento de o que automatizar primeiro.

Essa prática recomendada aplica-se não só a eventos de segurança gerados a partir de mensagens de log que representam atividades do usuário ou eventos de rede, como também a alterações detectadas na própria infraestrutura. A capacidade de detectar alterações, determinar se uma alteração foi apropriada e, em seguida, rotear essas informações para o fluxo de trabalho de correção correto é essencial para manter e validar uma arquitetura segura, no contexto de alterações em que a natureza de sua indesejabilidade é suficientemente sutil para que sua execução não possa ser impedida com uma combinação de configuração do AWS Identity and Access Management(IAM) e do AWS Organizations.

O Amazon GuardDuty e o AWS Security Hub fornecem mecanismos de agregação, desduplicação e análise para registros de log que também são disponibilizados a você por meio de outros serviços da AWS. O GuardDuty ingere, agrega e analisa informações de fontes como gerenciamento e eventos de dados do AWS CloudTrail, logs de DNS de VPC e logs de fluxo de VPC. O Security Hub pode ingerir, agregar e analisar a saída do GuardDuty AWS Config, do Amazon Inspector, Amazon Macie, do AWS Firewall Manager e de um número significativo de produtos de segurança de terceiros

disponíveis no AWS Marketplace e, se criado adequadamente, no seu próprio código. Tanto o GuardDuty quanto o Security Hub têm um modelo de membro administrador que pode agregar descobertas e insights em várias contas. O Security Hub geralmente é usado por clientes que têm um SIEM on-premises como um log do lado da AWS e um pré-processador e agregador de logs e alertas nos quais eles podem consumir o Amazon EventBridge por meio de um processador e encaminhador com base no AWS Lambda.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Avaliar os recursos de processamento de log: avalie as opções disponíveis para o processamento de logs.
  - Use Amazon OpenSearch Service to log and monitor (almost) everything (Usar o Amazon OpenSearch Service para registrar e monitorar (quase) tudo)
  - Encontre um parceiro especializado em soluções de registro e monitoramento
- Para começar a analisar logs do CloudTrail, experimente o Amazon Athena.
  - Como configurar o Athena para analisar logs do CloudTrail
- Implementar o login centralizado na AWS: consulte a solução de exemplo da AWS a seguir para centralizar o log de várias fontes.
  - Centralizar a solução de registro em log
- Implementar o registro em log centralizado com o parceiro: os parceiros da APN têm soluções para ajudar você a analisar os logs de forma centralizada.
  - Registro em log e monitoramento

Recursos

Documentos relacionados:

- AWS Answers: registro em log centralizado
- AWS Security Hub
- Amazon CloudWatch
- Amazon EventBridge
- Conceitos básicos: Amazon CloudWatch Logs
- Soluções de segurança parceiros: registro em log e monitoramento

Vídeos relacionados:

- Centrally Monitoring Resource Configuration and Compliance (Monitoramento centralizado de configuração e conformidade de recursos)
- Remediating Amazon GuardDuty and AWS Security Hub Findings (Correção do Amazon GuardDuty e descobertas do AWS Security Hub)
- Threat management in the cloud: Amazon GuardDuty and AWS Security Hub (Gerenciamento de ameaças na nuvem: Amazon GuardDuty e AWS Security Hub)

SEC04-BP03 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.

Na AWS, a investigação de eventos de interesse e informações sobre alterações potencialmente inesperadas em um fluxo de trabalho automatizado pode ser obtida com o Amazon EventBridge. Esse serviço fornece um mecanismo de regras escalável, projetado para processar formatos de eventos da AWS nativos (como eventos do AWS CloudTrail) e personalizados, que você pode gerar com base em sua aplicação. O Amazon GuardDuty também permite rotear eventos em um sistema de fluxo de trabalho para usuários que criam sistemas de resposta a incidentes (AWS Step Functions), uma conta de segurança central ou um bucket para análise posterior.

A detecção de alterações e o roteamento dessas informações para o fluxo de trabalho correto podem ser realizados com o uso do Regras do AWS Config e de pacotes de conformidade. O AWS Config detecta alterações nos serviços em escopo (embora com maior latência do que o EventBridge) e gera eventos que podem ser analisados usando o Regras do AWS Config para reversão, aplicação da política de conformidade e encaminhamento de informações aos sistemas, como plataformas de gerenciamento de alterações e sistemas operacionais de emissão de tíquetes. Além de escrever suas próprias funções do Lambda para responder a eventos do AWS Config, você também pode aproveitar o kit de desenvolvimento do Regras do AWS Confige uma biblioteca de código aberto do Regras do AWS Config. Os pacotes de conformidade são uma coleção de ações de correção e do Regras do AWS Config que você implanta como uma única entidade criada como um modelo YAML. O modelo de pacote de conformidade de amostra está disponível no pilar Segurança do Well-Architected.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Implementar alertas automatizados com o GuardDuty: o GuardDuty é um serviço de detecção de ameaças que monitora continuamente atividades mal-intencionadas e comportamentos não autorizados para proteger suas workloads e Contas da AWS. Habilite o GuardDuty e configure alertas automatizados.
- Automatizar o processo de investigação: desenvolva processos automatizados que investigam um evento e relatam informações a um administrador para economizar tempo.
  - Laboratório: Amazon GuardDuty na prática

Recursos

Documentos relacionados:

- AWS Answers: registro em log centralizado
- AWS Security Hub
- Amazon CloudWatch
- Amazon EventBridge
- Conceitos básicos: Amazon CloudWatch Logs
- Soluções de segurança parceiros: registro em log e monitoramento
- Como configurar o Amazon GuardDuty

Vídeos relacionados:

- Centrally Monitoring Resource Configuration and Compliance (Monitoramento centralizado de configuração e conformidade de recursos)
- Remediating Amazon GuardDuty and AWS Security Hub Findings (Correção do Amazon GuardDuty e descobertas do AWS Security Hub)
- Threat management in the cloud: Amazon GuardDuty and AWS Security Hub (Gerenciamento de ameaças na nuvem: Amazon GuardDuty e AWS Security Hub)

Exemplos relacionados:

- Laboratório: Implantação automatizada de controles de detecção

SEC04-BP04 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. Para cada mecanismo de detecção existente, você também deve ter um processo, na forma de um runbook ou playbook, para investigar. Por exemplo, quando você habilita o Amazon GuardDuty, ele gera diferentes descobertas. Você deve ter uma entrada de runbook para cada tipo de descoberta, por exemplo, se um cavalo de Troia for descoberto, seu runbook terá instruções simples que instruem alguém a investigar e corrigir o problema.

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

Orientação de implementação

- Descubra as métricas disponíveis para serviços da AWS: descubra as métricas disponíveis por meio do Amazon CloudWatch para os serviços que você está usando.
  - Documentação do serviço da AWS
  - Uso de métricas do Amazon CloudWatch
- Configure os alarmes do Amazon CloudWatch.
  - Como usar os alarmes do Amazon CloudWatch

Recursos

Documentos relacionados:

- Amazon CloudWatch
- Amazon EventBridge
- Soluções de segurança parceiros: registro em log e monitoramento

Vídeos relacionados:

- Centrally Monitoring Resource Configuration and Compliance (Monitoramento centralizado de configuração e conformidade de recursos)
- Remediating Amazon GuardDuty and AWS Security Hub Findings (Correção do Amazon GuardDuty e descobertas do AWS Security Hub)
- Threat management in the cloud: Amazon GuardDuty and AWS Security Hub (Gerenciamento de ameaças na nuvem: Amazon GuardDuty e AWS Security Hub)

Proteção de infraestrutura

Perguntas

- SEC 5 Como você protege seus recursos de rede?
- SEC 6 Como você protege seus recursos de computação?

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.

Práticas recomendadas

- SEC05-BP01 Criar camadas de rede
- SEC05-BP02 Controlar tráfego de todas as camadas
- SEC05-BP03 Automatizar a proteção da rede:
- SEC05-BP04 Implementar inspeção e proteção

SEC05-BP01 Criar camadas de rede

Agrupe componentes que compartilham requisitos de acessibilidade em camadas. Por exemplo, um cluster de banco de dados em uma nuvem privada virtual (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.

Os componentes como instâncias do Amazon Elastic Compute Cloud (Amazon EC2), clusters de banco de dados do Amazon Relational Database Service (Amazon RDS) e funções do AWS Lambda que compartilham requisitos de acessibilidade podem ser segmentados em camadas formadas por sub-redes. Por exemplo, um cluster de banco de dados do Amazon RDS em uma VPC sem necessidade de acesso à Internet deve ser colocado em sub-redes sem nenhuma rota para/da Internet. Essa abordagem em camadas para os controles reduz o impacto da configuração incorreta de uma única camada, o que pode permitir o acesso não intencional. Para o Lambda, você pode executar as funções em sua VPC para avançar os controles baseados em VPC.

Para uma conectividade de rede que possa incluir milhares de VPCs, contas da AWS e redes on-premises, você deve usar o AWS Transit Gateway. Ele atua como um hub que controla como o

tráfego é roteado entre todas as redes conectadas, que atuam como spokes. O tráfego entre uma Amazon Virtual Private Cloud e o AWS Transit Gateway permanece na rede privada da AWS, o que reduz vetores de ameaças externas, como ataques de negação de serviço distribuída (DDoS) e ameaças comuns, como injeção de SQL, cross-site scripting, falsificação de solicitações entre sites ou abuso de código de autenticação violado. O emparelhamento entre regiões do AWS Transit Gateway também criptografa o tráfego entre regiões sem um ponto único de falha ou gargalo de largura de banda.

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

Orientação de implementação

- Crie sub-redes na VPC: crie sub-redes para cada camada (em grupos que incluem várias zonas de disponibilidade) e associe tabelas de rotas para controlar o roteamento.
  - VPCs e sub-redes
  - Tabelas de rotas

Recursos

Documentos relacionados:

- AWS Firewall Manager
- Amazon Inspector
- Segurança da Amazon VPC
- Conceitos básicos do AWS WAF

Vídeos relacionados:

- AWS Transit Gateway reference architectures for many VPCs (Arquiteturas de referência do AWS Transit Gateway para várias VPCs)
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield (Aceleração e proteção de aplicações com o Amazon CloudFront, o AWS WAF e o AWS Shield)

Exemplos relacionados:

- Laboratório: Implantação automatizada da VPC

SEC05-BP02 Controlar tráfego de todas as camadas

ao projetar sua topologia de rede, você deve examinar os requisitos de conectividade de cada componente. Por exemplo, se um componente precisa de acesso à Internet (entrada e saída), conectividade com VPCs, serviços de borda e datacenters externos.

Uma VPC permite que você defina a topologia de rede que abrange uma Região da AWS com um intervalo de endereços IPv4 privados que você define ou um intervalo de endereços IPv6 que a AWS seleciona. Você deve aplicar vários controles com uma abordagem detalhada de defesa para tráfego de entrada e saída, incluindo o uso de grupos de segurança (firewall de inspeção com estado), Network ACLs, sub-redes e tabelas de rotas. Você pode criar sub-redes em uma zona de disponibilidade dentro de uma VPC. Cada sub-rede tem uma tabela de rotas associada que define regras de roteamento para gerenciar os caminhos do tráfego dentro da sub-rede. Você pode definir uma sub-rede roteável na Internet com uma rota que siga até um gateway da Internet ou gateway NAT associado à VPC ou que passe por outra VPC.

Quando uma instância, um banco de dados do Amazon Relational Database Service(Amazon RDS) ou outro serviço é executado em uma VPC, ela tem seu próprio grupo de segurança por interface de rede. Esse firewall está fora da camada do sistema operacional e pode ser usado para definir regras para o tráfego permitido de entrada e saída. Você também pode definir relacionamentos entre grupos de segurança. Por exemplo, as instâncias em um grupo de segurança no nível do banco de dados aceitam somente o tráfego de instâncias no nível do aplicativo, por referência aos grupos de segurança aplicados às instâncias envolvidas. A menos que você esteja usando protocolos não baseados em TCP, não deve ser necessário ter uma instância do Amazon Elastic Compute Cloud(Amazon EC2) diretamente acessível pela Internet (mesmo com portas restritas por grupos de segurança) sem um balanceador de carga ou o CloudFront. Isso ajuda a protegê-lo contra acesso não intencional surgido por um problema de sistema operacional ou aplicativo. Uma sub-rede também pode ter uma Network ACL anexada a ela, que atua como um firewall sem estado. Você deve configurar a Network ACL para restringir a abrangência do tráfego permitido entre camadas. Observe que é preciso definir regras de entrada e de saída.

Alguns serviços da AWS requerem componentes para acessar a Internet para fazer chamadas de API, onde os endpoints de API da AWS estão localizados. Outros serviços da AWS usam VPC endpoints dentro das suas Amazon VPCs. Muitos serviços da AWS, incluindo o Amazon S3 e o Amazon DynamoDB, oferecem suporte a endpoints da VPC, e essa tecnologia foi generalizada no AWS PrivateLink. Recomendamos o uso dessa abordagem para acessar serviços da AWS, serviços de terceiros e seus próprios serviços hospedados em outras VPCs com segurança. Todo o tráfego de rede do AWS PrivateLink permanece no backbone global da AWS e nunca atravessa

a Internet. A conectividade só pode ser iniciada pelo consumidor do serviço e não pelo provedor do serviço. O uso do AWS PrivateLink para acesso a serviços externos permite criar VPCs air-gapped sem acesso à Internet e ajuda a proteger suas VPCs de vetores de ameaças externas. Os serviços de terceiros podem usar o AWS PrivateLink para permitir que os clientes se conectem aos serviços de suas VPCs por meio de endereços IP privados. Para ativos da VPC que precisam estabelecer conexões de saída com a Internet, elas podem ser feitas somente de saída (unidirecional) por meio de um gateway NAT gerenciado pela AWS, de um gateway da Internet somente de saída ou de proxies de Web criados e gerenciados por você.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Controlar o tráfego de rede em uma VPC: implemente as práticas recomendadas de VPC para controlar o tráfego.
  - Segurança da Amazon VPC
  - VPC endpoints
  - Grupo de segurança da Amazon VPC
  - ACLs de rede
- Controlar o tráfego na borda: implemente serviços de borda, como o Amazon CloudFront, para fornecer uma camada adicional de proteção e outros recursos.
  - Casos de uso do Amazon CloudFront
  - AWS Global Accelerator
  - AWS Web Application Firewall (AWS WAF)
  - Amazon Route 53
  - Roteamento de entrada da Amazon VPC
- Controlar o tráfego de rede privada: implemente serviços que protegem o tráfego privado da sua workload.
  - Emparelhamento de Amazon VPC
  - Serviços de endpoint da Amazon VPC (AWS PrivateLink)
  - Amazon VPC Transit Gateway
  - AWS Direct Connect
  - AWS Site-to-Site VPN
  - AWS Client VPN

- Pontos de acesso do Amazon S3

Recursos

Documentos relacionados:

- AWS Firewall Manager
- Amazon Inspector
- Conceitos básicos do AWS WAF

Vídeos relacionados:

- AWS Transit Gateway reference architectures for many VPCs (Arquiteturas de referência do AWS Transit Gateway para várias VPCs)
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield (Aceleração e proteção de aplicações com o Amazon CloudFront, o AWS WAF e o AWS Shield)

Exemplos relacionados:

- Laboratório: Implantação automatizada da VPC

SEC05-BP03 Automatizar a proteção da 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 às ameaças atuais e reduzir seu impacto. Um firewall de aplicação Web é um exemplo de onde você pode automatizar a proteção de rede; por exemplo, usando a solução AWS WAF Security Automations (https://github.com/awslabs/aws-waf-security-automations) para bloquear automaticamente solicitações originadas de endereços IP associados a agentes de ameaças conhecidos.

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

Orientação de implementação

- Automatize a proteção para tráfego baseado na Web: a AWS oferece uma solução que usa o AWS CloudFormation para implantar automaticamente um conjunto de regras do AWS WAF projetadas

para filtrar ataques comuns baseados na Web. Os usuários podem selecionar entre recursos de proteção pré-configurados que definem as regras incluídas em uma lista de controle de acesso da Web (ACL da Web) do AWS WAF.

  - Automações de segurança do AWS WAF
- Considere as soluções de AWS Partner: os parceiros da AWS oferecem centenas de produtos líderes do setor que são equivalentes, idênticos ou se integram aos controles existentes nos seus ambientes on-premises. Esses produtos complementam os serviços da AWS já existentes para que os clientes possam implantar uma arquitetura de segurança abrangente e obter uma experiência mais uniforme na nuvem e no ambiente on-premises.
  - Segurança da infraestrutura

Recursos

Documentos relacionados:

- AWS Firewall Manager
- Amazon Inspector
- Segurança da Amazon VPC
- Conceitos básicos do AWS WAF

Vídeos relacionados:

- AWS Transit Gateway reference architectures for many VPCs (Arquiteturas de referência do AWS Transit Gateway para várias VPCs)
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield (Aceleração e proteção de aplicações com o Amazon CloudFront, o AWS WAF e o AWS Shield)

Exemplos relacionados:

- Laboratório: Implantação automatizada da VPC

SEC05-BP04 Implementar inspeção e proteção

Inspecione e filtre o tráfego em cada camada. É possível inspecionar suas configurações de VPC quanto a possíveis acessos não intencionais usando o VPC Network Access Analyzer. Especifique seus requisitos de acesso à rede e identifique possíveis caminhos de rede que não os atendem.

Para componentes que fazem transações por meio de protocolos baseados em HTTP, um firewall de aplicativo Web pode ajudar a proteger contra ataques comuns. AWS WAF é um firewall para aplicativos web que permite monitorar e bloquear solicitações HTTP(s) que correspondem às regras configuráveis que são encaminhadas para uma API do Amazon API Gateway, o Amazon CloudFront ou um Application Load Balancer. Para começar a usar o AWS WAF, você pode usar o AWS Managed Rules em combinação com as suas próprias ou usar integrações de parceiros existentes.

Para gerenciar o AWS WAF, proteções do AWS Shield Advanced e grupos de segurança do Amazon VPC no AWS Organizations, você pode usar o AWS Firewall Manager. Ele permite configurar e gerenciar centralmente regras de firewall entre contas e aplicativos, simplificando a imposição de regras comuns em escala. Ele também permite que você responda rapidamente a ataques, usando o AWS Shield Advancedou soluções capazes de bloquear automaticamente solicitações indesejadas para suas aplicações Web. O Firewall Manager também funciona com o AWS Network Firewall. O AWS Network Firewall é um serviço gerenciado que usa um mecanismo de regras para fornecer controle refinado sobre o tráfego de rede com e sem estado. Ele oferece suporte às especificações do sistema de prevenção de intrusões (IPS) de código aberto compatível com Suricata para regras para ajudar a proteger sua workload.

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

Orientação de implementação

- Configure o Amazon GuardDuty: o GuardDuty é um serviço de detecção de ameaças que monitora continuamente atividades mal-intencionadas e comportamentos não autorizados para proteger suas workloads e Contas da AWS. Habilite o GuardDuty e configure alertas automatizados.
  - Amazon GuardDuty
  - Laboratório: Implantação automatizada de controles de detecção
- Configure os logs de fluxo da nuvem privada virtual (VPC): os logs de fluxo da VPC é um recurso que permite capturar informações sobre o tráfego de IP direcionado e proveniente de interfaces de rede na sua VPC. Os dados de log de fluxo podem ser publicados no Amazon CloudWatch Logs e no Amazon Simple Storage Service (Amazon S3). Depois de criar um log de fluxo, você pode recuperar e visualizar seus dados no destino escolhido.
- Considere o espelhamento de tráfego da VPC: o espelhamento de tráfego é um recurso da Amazon VPC que pode ser usado para copiar o tráfego de rede de uma interface de rede elástica de instâncias do Amazon Elastic Compute Cloud (Amazon EC2) e enviá-lo para dispositivos de segurança e monitoramento fora de banda para inspeção de conteúdo, monitoramento de ameaças e solução de problemas.

- Espelhamento de tráfego de VPC

Recursos

Documentos relacionados:

- AWS Firewall Manager
- Amazon Inspector
- Segurança da Amazon VPC
- Conceitos básicos do AWS WAF

Vídeos relacionados:

- AWS Transit Gateway reference architectures for many VPCs (Arquiteturas de referência do AWS Transit Gateway para várias VPCs)
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield (Aceleração e proteção de aplicações com o Amazon CloudFront, o AWS WAF e o AWS Shield)

Exemplos relacionados:

- Laboratório: Implantação automatizada da VPC

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. Recursos de computação incluem instâncias do EC2, contêineres, funções do AWS Lambda, serviços de banco de dados, dispositivos de IoT e muito mais.

Práticas recomendadas

- SEC06-BP01 Fazer o gerenciamento de vulnerabilidades
- SEC06-BP02 Reduzir a superfície de ataque
- SEC06-BP03 Implementar serviços gerenciados
- SEC06-BP04 Automatizar a proteção da computação
- SEC06-BP05 Permitir que as pessoas executem ações a uma distância
- SEC06-BP06 Validar a integridade do software

SEC06-BP01 Fazer o gerenciamento de vulnerabilidades

Verifique e corrija com frequência vulnerabilidades no código, nas dependências e na infraestrutura para proteger-se contra novas ameaças.

Começando com a configuração de sua infraestrutura de computação, é possível automatizar a criação e atualização de recursos usando o AWS CloudFormation. O CloudFormation permite criar modelos escritos em YAML ou JSON, usando exemplos da AWS ou escrevendo os seus próprios. Isso permite criar modelos de infraestrutura seguros por padrão que você pode verificar com o CloudFormation Guard, para economizar tempo e reduzir o risco de erros de configuração. Você pode criar a infraestrutura e implantar suas aplicações usando entrega contínua; por exemplo, com o AWS CodePipeline, para automatizar a criação, o teste e o lançamento.

Você é responsável pelo gerenciamento de patches para seus recursos do AWS, incluindo instâncias do Amazon Elastic Compute Cloud(Amazon EC2), imagens de máquina da Amazon (AMIs) e muitos outros recursos de computação. Para instâncias do Amazon EC2, o Patch Manager do AWS Systems Manager automatiza o processo de aplicação de patches em instâncias gerenciadas com atualizações relacionadas à segurança e com outros tipos de atualizações. Você pode usar o gerenciador de patches para aplicar patches a sistemas operacionais e aplicações. (No Windows Server, o suporte à aplicação é limitado a atualizações para aplicações da Microsoft.) Use o Patch Manager para instalar pacotes de serviços em instâncias do Windows e realizar atualizações de versões secundárias em instâncias do Linux. Corrija frotas de instâncias do Amazon EC2 ou de seus servidores on-premises e máquinas virtuais (VMs) por tipo de sistema operacional. Isso inclui versões compatíveis do Windows Server, Amazon Linux, Amazon Linux 2, CentOS, Debian Server, Oracle Linux, Red Hat Enterprise Linux (RHEL), SUSE Linux Enterprise Server (SLES) e Ubuntu Server. Você pode verificar instâncias para ver apenas um relatório de patches ausentes ou verificar e instalar automaticamente todos os patches ausentes.

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

Orientação de implementação

- Configure o Amazon Inspector: o Amazon Inspector testa a acessibilidade de rede das instâncias do Amazon Elastic Compute Cloud (Amazon EC2) e o estado de segurança das aplicações executadas nessas instâncias. O Amazon Inspector avalia aplicações para exposição, vulnerabilidades e desvios das práticas recomendadas.
  - O que é o Amazon Inspector?
- Escaneie o código-fonte: escaneie bibliotecas e dependências em busca de vulnerabilidades.
  - Amazon CodeGuru

- OWASP: source code analysis tools

Recursos

Documentos relacionados:

- AWS Systems Manager
- Replacing a Bastion Host with Amazon EC2 Systems Manager (Como substituir um host traga a sua própria licença pelo Amazon EC2 Systems Manager)
- Security Overview of AWS Lambda (Visão geral de segurança do AWS Lambda)

Vídeos relacionados:

- Running high-security workloads on Amazon EKS (Execução de workloads de alta segurança no Amazon EKS)
- Securing Serverless and Container Services (Proteção de serviços com tecnologia sem servidor e de contêiner)
- Security best practices for the Amazon EC2 instance metadata service (Práticas recomendadas de segurança para o serviço de metadados de instância do Amazon EC2

Exemplos relacionados:

- Laboratório: Implantação automatizada do firewall de aplicações Web

SEC06-BP02 Reduzir a superfície de ataque

Reduza a exposição ao acesso não intencional protegendo os sistemas operacionais e minimizando componentes, bibliotecas e serviços consumíveis externamente em uso. Primeiro, diminua o número de componentes não utilizados, sejam eles pacotes de sistema operacional ou aplicações para workloads baseadas no Amazon Elastic Compute Cloud (Amazon EC2), sejam eles módulos de software externos no código, para todas as workloads. Encontre muitos guias de configuração de proteção e segurança para sistemas operacionais comuns e software de servidor. Por exemplo, você pode começar com o Center for Internet Security e iterar.

No Amazon EC2, e possível criar as próprias imagens de máquina da Amazon (AMIs), corrigidas e reforçadas, para ajudar você a atender aos requisitos de segurança específicos da sua organização. Os patches e outros controles de segurança aplicados na AMI são efetivos no momento em que

foram criados. Eles não são dinâmicos, a menos que você modifique após a inicialização, por exemplo, com o AWS Systems Manager.

É possível simplificar o processo de criação de AMIs seguras com o EC2 Image Builder. O EC2 Image Builder reduz significativamente o esforço necessário para criar e manter imagens douradas sem escrever e manter a automação. Quando as atualizações de software ficam disponíveis, o Image Builder produz automaticamente uma nova imagem sem exigir que os usuários iniciem manualmente as compilações de imagem. O EC2 Image Builder permite validar facilmente a funcionalidade e a segurança de suas imagens antes de usá-las na produção com testes fornecidos pela AWS e seus próprios testes. Também é possível aplicar as configurações de segurança fornecidas pela AWS para proteger ainda mais suas imagens para atender aos critérios de segurança internos. Por exemplo, você pode produzir imagens em conformidade com o padrão do Guia de implementação técnica de segurança (STIG) usando modelos fornecidos pela AWS.

Com ferramentas de análise de código estático de terceiros é possível identificar problemas de segurança comuns, como limites de entrada de função não verificados, bem como vulnerabilidades e exposições comuns (CVEs) aplicáveis. Você pode usar o Amazon CodeGuru para os idiomas compatíveis. As ferramentas de verificação de dependência também podem ser usadas para determinar se as bibliotecas com as quais o código está vinculado são as versões mais recentes, estão livres de CVEs e têm condições de licenciamento que atendem aos requisitos da política de software.

Usando o Amazon Inspector, você pode executar avaliações de configuração de CVEs conhecidas em suas instâncias, avaliar parâmetros de segurança e automatizar a notificação de defeitos. O Amazon Inspector é executado em instâncias de produção ou em um pipeline de compilação e notifica desenvolvedores e engenheiros quando descobertas estão presentes. Você pode acessar as descobertas programaticamente e direcionar sua equipe para os registros em atraso e os sistemas de rastreamento de bugs. EC2 Image Builder pode ser usado para manter imagens de servidor (AMIs) com aplicação automática de patches, aplicação de políticas de segurança fornecidas pela AWS e outras personalizações. Ao usar contêineres, implemente a Verificação de imagens do ECR no pipeline de compilação e regularmente no repositório de imagens para procurar CVEs nos contêineres.

Embora o Amazon Inspector e outras ferramentas sejam eficazes na identificação de configurações e CVEs presentes, outros métodos são necessários para testar a carga de trabalho no nível do aplicativo. Fuzzing é um método conhecido de encontrar erros usando automação para injetar dados malformados em campos de entrada e outras áreas do aplicativo.

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

Orientação de implementação

- Configure os sistemas operacionais: configure os sistemas operacionais para atender às práticas recomendadas.
  - Securing Amazon Linux
  - Securing Microsoft Windows Server
- Configure recursos em contêiner para atender às práticas recomendadas de segurança.
- Implemente as práticas recomendadas do AWS Lambda.
  - Práticas recomendadas do AWS Lambda

Recursos

Documentos relacionados:

- AWS Systems Manager
- Replacing a Bastion Host with Amazon EC2 Systems Manager (Como substituir um host traga a sua própria licença pelo Amazon EC2 Systems Manager)
- Security Overview of AWS Lambda (Visão geral de segurança do AWS Lambda)

Vídeos relacionados:

- Running high-security workloads on Amazon EKS (Execução de workloads de alta segurança no Amazon EKS)
- Securing Serverless and Container Services (Proteção de serviços com tecnologia sem servidor e de contêiner)
- Security best practices for the Amazon EC2 instance metadata service (Práticas recomendadas de segurança para o serviço de metadados de instância do Amazon EC2

Exemplos relacionados:

- Laboratório: Implantação automatizada do firewall de aplicações Web

SEC06-BP03 Implementar serviços gerenciados

Implemente serviços que gerenciam recursos, como o Amazon Relational Database Service (Amazon RDS), o AWS Lambda e o Amazon Elastic Container Service (Amazon ECS), para reduzir

as tarefas de manutenção de segurança como parte do modelo de responsabilidade compartilhada. Por exemplo, o Amazon RDS ajuda você a configurar, operar e escalar um banco de dados relacional, automatiza tarefas de administração, como provisionamento de hardware, configuração de banco de dados, aplicação de patches e backups. Isso significa que você tem mais tempo livre para se concentrar na proteção da aplicação de outras maneiras descritas no AWS Well-Architected Framework. O Lambda permite executar código sem provisionar nem gerenciar servidores e, portanto, você só precisa se concentrar na conectividade, na invocação e na segurança em nível de código, e não na infraestrutura ou no sistema operacional.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Explorar os serviços disponíveis: explore, teste e implemente serviços que gerenciam recursos, como Amazon RDS, AWS Lambda e Amazon ECS.

Recursos

Documentos relacionados:

- Site da AWS
- AWS Systems Manager
- Replacing a Bastion Host with Amazon EC2 Systems Manager (Como substituir um bastion host com o Amazon EC2 Systems Manager)
- Security Overview of AWS Lambda (Visão geral de segurança do AWS Lambda)

Vídeos relacionados:

- Running high-security workloads on Amazon EKS (Execução de workloads de alta segurança no Amazon EKS)
- Securing Serverless and Container Services (Proteção de serviços com tecnologia sem servidor e de contêiner)
- Security best practices for the Amazon EC2 instance metadata service (Práticas recomendadas de segurança para o serviço de metadados de instância do Amazon EC2)

Exemplos relacionados:

- Laboratório: AWS Certificate Manager Request Public Certificate

SEC06-BP04 Automatizar a proteção da computação

Automatize seus mecanismos de computação de proteção, incluindo gerenciamento de vulnerabilidades, redução da superfície de ataque e gerenciamento de recursos. A automação ajudará você a investir tempo para proteger outros aspectos da carga de trabalho e reduzir o risco de erros humanos.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Automatizar o gerenciamento de configuração: aplique e valide configurações seguras automaticamente usando uma ferramenta ou um serviço de gerenciamento de configuração.
  - AWS Systems Manager
  - AWS CloudFormation
  - Laboratório: Implantação automatizada da VPC
  - Laboratório: Implantação automatizada da aplicação Web no EC2

- Automatizar a aplicação de patches para instâncias do Amazon Elastic Compute Cloud(Amazon EC2): o Patch Manager do AWS Systems Manager automatiza o processo de aplicação de patches em instâncias gerenciadas com atualizações relacionadas à segurança e com outros tipos de atualizações. Você pode usar o gerenciador de patches para aplicar patches a sistemas operacionais e aplicações.
  - AWS Systems Manager Patch Manager
  - Correção centralizada de várias contas e várias regiões com automação do AWS Systems Manager

- Implementar detecção e prevenção de intrusão: implemente uma ferramenta de detecção e prevenção de invasões para monitorar e interromper atividades maliciosas nas instâncias.
- Considerar as soluções de AWS Partner: os parceiros da AWS oferecem centenas de produtos líderes do setor que são equivalentes, idênticos ou se integram aos controles existentes nos seus ambientes on-premises. Esses produtos complementam os serviços da AWS já existentes para que os clientes possam implantar uma arquitetura de segurança abrangente e obter uma experiência mais uniforme na nuvem e no ambiente on-premises.

- Segurança da infraestrutura

Recursos

Documentos relacionados:

- AWS CloudFormation
- AWS Systems Manager
- AWS Systems Manager Patch Manager
- Correção centralizada de várias contas e várias regiões com automação do AWS Systems Manager
- Segurança da infraestrutura
- Replacing a Bastion Host with Amazon EC2 Systems Manager (Como substituir um bastion host com o Amazon EC2 Systems Manager)
- Security Overview of AWS Lambda (Visão geral de segurança do AWS Lambda)

Vídeos relacionados:

- Running high-security workloads on Amazon EKS (Execução de workloads de alta segurança no Amazon EKS)
- Securing Serverless and Container Services (Proteção de serviços com tecnologia sem servidor e de contêiner)
- Security best practices for the Amazon EC2 instance metadata service (Práticas recomendadas de segurança para o serviço de metadados de instância do Amazon EC2)

Exemplos relacionados:

- Laboratório: Implantação automatizada do firewall de aplicações Web
- Laboratório: Implantação automatizada da aplicação Web no EC2

SEC06-BP05 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 Amazon Elastic Compute Cloud (Amazon EC2) usando

infraestruturas como código e gerenciar instâncias do Amazon EC2 com ferramentas, como o AWS Systems Manager, em vez de permitir acesso direto, ou por meio de um host traga a sua própria licença. O AWS Systems Managerpode automatizar uma variedade de tarefas de manutenção e implantação, usando recursos que incluem fluxos de trabalho de automação ,, documentos (playbooks) e o Run Command. O AWS CloudFormation empilha a compilação com base em pipelines e pode automatizar tarefas de implantação e gerenciamento de infraestrutura sem usar diretamente o AWS Management Console ou APIs.

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

Orientação de implementação

- Substitua o acesso ao controle: substitua o acesso ao console (SSH ou RDP) a instâncias com o Run Command do AWS Systems Manager para automatizar tarefas de gerenciamento.

- AWS Systems Manager Run Command

Recursos

Documentos relacionados:

- AWS Systems Manager
- AWS Systems Manager Run Command
- Replacing a Bastion Host with Amazon EC2 Systems Manager (Como substituir um host traga a sua própria licença pelo Amazon EC2 Systems Manager)
- Security Overview of AWS Lambda (Visão geral de segurança do AWS Lambda)

Vídeos relacionados:

- Running high-security workloads on Amazon EKS (Execução de workloads de alta segurança no Amazon EKS)
- Securing Serverless and Container Services (Proteção de serviços com tecnologia sem servidor e de contêiner)
- Security best practices for the Amazon EC2 instance metadata service (Práticas recomendadas de segurança para o serviço de metadados de instância do Amazon EC2

Exemplos relacionados:

- Laboratório: Implantação automatizada do firewall de aplicações Web

SEC06-BP06 Validar a integridade do software

Implemente mecanismos (por exemplo, assinatura de código) para validar se o software, o código e as bibliotecas usados na workload são de fontes confiáveis e não foram adulterados. Por exemplo, você deve verificar o certificado de assinatura de código de binários e scripts para confirmar o autor e garantir que ele não tenha sido adulterado desde que foi criado pelo autor. AWS Signer pode ajudar a garantir a confiança e a integridade do código gerenciando centralmente o ciclo de vida de assinatura de código, incluindo certificação de assinatura e chaves públicas e privadas. Você pode aprender a usar padrões avançados e práticas recomendadas para assinatura de código com o AWS Lambda. Além disso, uma soma de verificação do software que você faz download, em comparação com a soma de verificação do provedor, pode ajudar a garantir que ela não tenha sido adulterada.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Investigar os mecanismo: a assinatura de código é um mecanismo que pode ser usado para validar a integridade do software.
  - NIST: Considerações de segurança para assinatura de código

Recursos

Documentos relacionados:

- AWS Signer
- New – Code Signing, a Trust and Integrity Control for AWS Lambda (Novo: assinatura de código, um controle de confiança e integridade para o AWS Lambda)

Proteção de dados

Perguntas

- SEC 7 Como você classifica seus dados?
- SEC 8 Como você protege seus dados em repouso?
- SEC 9 Como você protege seus dados em trânsito?

SEC 7 Como você classifica seus 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.

Práticas recomendadas

- SEC07-BP01 Identificar os dados em sua workload
- SEC07-BP02 Definir controles de proteção de dados
- SEC07-BP03 Automatizar a identificação e a classificação
- SEC07-BP04 Definir o gerenciamento do ciclo de vida de dados

SEC07-BP01 Identificar os dados em sua workload

você precisa conhecer o tipo e a classe dos dados processados pela carga de trabalho, os processos de negócios associados, o proprietário dos dados, os requisitos legais e de conformidade aplicáveis, onde estão 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 Personally Identifiable Information (PII – Informações de identificação pessoal) do cliente ou se os dados são para acesso mais restrito, como propriedade intelectual, dados legalmente privilegiados ou marcados como confidenciais, e muito mais. Ao gerenciar cuidadosamente um sistema de classificação de dados apropriado, juntamente com o nível de requisitos de proteção de cada workload, é possível mapear os controles e o nível de acesso ou proteção apropriados aos dados. Por exemplo, o conteúdo voltado para o público está disponível para qualquer pessoa acessar, enquanto o conteúdo importante é criptografado e armazenado de maneira protegida que requer acesso autorizado a uma chave para descriptografá-lo.

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

Orientação de implementação

- Considere descobrir dados usando o Macie: o Amazon Macie reconhece dados confidenciais, como informações de identificação pessoal (PII) ou propriedade intelectual.
  - Amazon Macie

Recursos

Documentos relacionados:

- Amazon Macie
- Whitepaper Classificação de dados
- Conceitos básicos do Amazon Macie

Vídeos relacionados:

- Introducing the New Amazon Macie (Apresentação do novo Amazon Macie)

SEC07-BP02 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.

Usando tags de recursos, separar contas da AWS por confidencialidade (e potencialmente também por advertência, enclave ou comunidade de interesse), políticas do IAM, SCPs do AWS Organizations, AWS Key Management Service (AWS KMS) e AWS CloudHSM, você pode definir e implementar as políticas de classificação e proteção de dados com criptografia. Por exemplo, se você tiver buckets do S3 que contêm dados altamente críticos ou instâncias do Amazon Elastic Compute Cloud (Amazon EC2) que processam dados confidenciais, eles poderão ser marcados com uma tag `Project=ABC` . Somente a equipe imediata sabe o que o código do projeto significa e fornece meios de usar o controle de acesso baseado em atributos. Você pode definir os níveis de acesso às chaves de criptografia do AWS KMS por meio de políticas de chave e concessões para garantir que somente os serviços apropriados tenham acesso ao conteúdo confidencial por meio de um mecanismo seguro. Se você estiver tomando decisões de autorização com base em tags, certifique-se de que as permissões nas tags sejam definidas adequadamente usando políticas de tags no AWS Organizations.

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

Orientação de implementação

- Defina o esquema de identificação e classificação de dados: a identificação e a classificação de seus dados são realizadas para avaliar o potencial impacto e o tipo de dados que você está armazenando e quem deve acessá-los.
  - Documentação da AWS

- Descubra os controles disponíveis da AWS: descubra os controles de segurança para os serviços da AWS que você usa ou planeja usar. Muitos serviços têm uma seção de segurança em sua documentação.
  - Documentação da AWS
- Identificar recursos de conformidade da AWS: identifique os recursos da AWS disponíveis para ajudar.
  - https://aws.amazon.com/compliance/

Recursos

Documentos relacionados:

- Documentação da AWS
- Whitepaper Classificação de dados
- Conceitos básicos do Amazon Macie
- Texto ausente

Vídeos relacionados:

- Introducing the New Amazon Macie (Apresentação do novo Amazon Macie)

SEC07-BP03 Automatizar a identificação e a classificação

automatizar a identificação e a classificação de dados pode ajudar a implementar os controles corretos. O uso de automação para isso, em vez de acesso direto de uma pessoa, reduz o risco de erros humanos e exposição. Você deve avaliar o uso de uma ferramenta, como o Amazon Macie, que usa machine learning para descobrir, classificar e proteger automaticamente dados confidenciais na AWS. O Amazon Macie reconhece dados confidenciais, como informações de identificação pessoal (PII) ou propriedade intelectual, e fornece painéis e alertas que dão visibilidade sobre como esses dados estão sendo acessados ou movidos.

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

Orientação de implementação

- Use o Amazon Simple Storage Service (Amazon S3) Inventory: o Amazon S3 Inventory é uma das ferramentas que você pode usar para auditar e gerar relatórios sobre o status de replicação e criptografia de seus objetos.
  - Amazon S3 Inventory
- Considere o Amazon Macie: O Amazon Macie usa o machine learning para descobrir e classificar automaticamente os dados armazenados no Amazon S3.
  - Amazon Macie

Recursos

Documentos relacionados:

- Amazon Macie
- Amazon S3 Inventory
- Whitepaper Classificação de dados
- Conceitos básicos do Amazon Macie

Vídeos relacionados:

- Introducing the New Amazon Macie (Apresentação do novo Amazon Macie)

SEC07-BP04 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 a duração pela qual você retém dados, processos de destruição de dados, gerenciamento de acesso a dados, transformação de dados e compartilhamento de dados devem ser considerados. Ao escolher uma metodologia de classificação de dados, equilibre usabilidade e acesso. Considere também os vários níveis de acesso e nuances para implementar uma abordagem segura, mas utilizável, para cada nível. Sempre use uma abordagem de defesa detalhada e reduza o acesso humano a dados e mecanismos para transformar, excluir ou copiar dados. Por exemplo, exija que os usuários se autentiquem fortemente em uma aplicação e conceda a ela, e não aos usuários, a permissão de acesso necessária para executar uma ação a distância. Além disso, garanta que os usuários venham de um caminho de rede confiável e exijam acesso às chaves de descriptografia. Use ferramentas como painéis ou relatórios

automatizados para fornecer aos usuários informações extraídas dos dados e não acesso direto aos dados.

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

Orientação de implementação

- Identificar tipos de dados: identifique os tipos de dados que você está armazenando ou processando em sua workload. Esses dados podem ser texto, imagens, bancos de dados binários, entre outros.

Recursos

Documentos relacionados:

- Whitepaper Classificação de dados
- Conceitos básicos do Amazon Macie

Vídeos relacionados:

- Introducing the New Amazon Macie (Apresentação do novo Amazon Macie)

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.

Práticas recomendadas

- SEC08-BP01 Implementar gerenciamento de chaves seguro
- SEC08-BP02 Aplicar criptografia em repouso
- SEC08-BP03 Automatizar a proteção de dados em repouso
- SEC08-BP04 Impor o controle de acesso
- SEC08-BP05 Usar mecanismos para evitar que as pessoas acessem os dados

SEC08-BP01 Implementar gerenciamento de chaves seguro

Ao definir uma abordagem de criptografia que inclui armazenamento, rotação e controle de acesso das chaves, você pode ajudar a proteger o conteúdo contra usuários não autorizados e contra exposição desnecessária a usuários autorizados. O AWS Key Management Service (AWS KMS) ajuda a gerenciar chaves de criptografia e se integra a vários serviços da AWS. Este serviço fornece armazenamento durável, seguro e redundante para as chaves do AWS KMS. Você pode definir seus alias principais e políticas de nível-chave. As políticas ajudam a definir os administradores de chaves e os usuários de chaves. Além disso, o AWS CloudHSM é um módulo de segurança de hardware baseado na nuvem (HSM) que permite gerar e usar facilmente suas próprias chaves de criptografia na Nuvem AWS. Ele ajuda a atender aos requisitos de conformidade corporativa, contratual e regulamentar para segurança de dados usando HSMs validados pelo FIPS 140-2 Nível 3.

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

Orientação de implementação

- Implemente o AWS KMS: o AWS KMS facilita a criação e o gerenciamento de chaves e o controle do uso de criptografia em uma ampla variedade de serviços da AWS e em aplicações. O AWS KMS é um serviço seguro e resiliente que usa módulos de segurança de hardware validados pelo FIPS 140-2 para proteger suas chaves.
  - Conceitos básicos: AWS Key Management Service (AWS KMS)
- Considere o SDK de criptografia da AWS: use o SDK de criptografia da AWS com a integração do AWS KMS quando sua aplicação precisar criptografar dados do lado do cliente.
  - SDK de criptografia da AWS

Recursos

Documentos relacionados:

- AWS Key Management Service
- Ferramentas e serviços criptográficos da AWS
- Conceitos básicos: AWS Key Management Service (AWS KMS)
- Proteção de dados usando criptografia do Amazon S3

Vídeos relacionados:

- How Encryption Works in AWS (Como a criptografia funciona no AWS Backup)
- Securing Your Block Storage on AWS (Como proteger o armazenamento em bloco na AWS)

SEC08-BP02 Aplicar criptografia em repouso

Garanta que a única maneira de armazenar dados seja usando a criptografia. O AWS Key Management Service (AWS KMS) se integra perfeitamente a muitos serviços da AWS para facilitar a criptografia de todos os seus dados em repouso. Por exemplo, no Amazon Simple Storage Service (Amazon S3), você pode definir a criptografia padrão em um bucket para que todos os novos objetos sejam criptografados automaticamente. Além disso, o Amazon Elastic Compute Cloud (Amazon EC2) e Amazon S3 oferecem suporte à imposição de criptografia ao definir a criptografia padrão. Você pode usar o Regras do AWS Config para verificar automaticamente se está usando criptografia, por exemplo, para volumes do Amazon Elastic Block Store (Amazon EBS), instâncias do Amazon Relational Database Service (Amazon RDS)e aos Amazon S3.

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

Orientação de implementação

- Impor criptografia em repouso para o Amazon Simple Storage Service (Amazon S3): implemente a criptografia padrão do bucket do Amazon S3.
  - Como habilito a criptografia padrão para um bucket do S3?
- Use o AWS Secrets Manager: o Secrets Manager é um serviço da AWS que facilita o gerenciamento de segredos. Segredos podem ser credenciais de banco de dados, senhas, chaves de API de terceiros e até texto arbitrário.
  - AWS Secrets Manager
- Configure a criptografia padrão para volumes do EBS: especifique que você deseja que todos os volumes do EBS recém-criados sejam criados em formato criptografado, com a opção de usar a chave padrão fornecida pela AWS ou uma chave que você criar.
  - Criptografia padrão para volumes do EBS
- Configurar imagens de máquina da Amazon (AMIs) criptografadas: a cópia de uma AMI existente com criptografia habilitada criptografará automaticamente os volumes raiz e os snapshots.
  - AMIs com snapshots criptografados
- Configurar a criptografia do Amazon Relational Database Service (Amazon RDS): configure a criptografia para seus clusters de banco de dados Amazon RDS e snapshots em repouso ativando a opção de criptografia.

  - Criptografia de recursos do Amazon RDS
- Configure a criptografia em serviços adicionais da AWS: para os serviços da AWS que você usa, determine os recursos de criptografia.
  - Documentação da AWS

Recursos

Documentos relacionados:

- AMIs com snapshots criptografados
- Ferramentas de criptografia da AWS
- Documentação da AWS
- SDK de criptografia da AWS
- Whitepaper de detalhes criptográficos do AWS KMS
- AWS Key Management Service
- AWS Secrets Manager
- Ferramentas e serviços criptográficos da AWS
- Criptografia do Amazon EBS
- Criptografia padrão para volumes do EBS
- Criptografia de recursos do Amazon RDS
- Como habilito a criptografia padrão para um bucket do S3?
- Proteção de dados usando criptografia do Amazon S3

Vídeos relacionados:

- How Encryption Works in AWS (Como a criptografia funciona no AWS Backup)
- Securing Your Block Storage on AWS (Como proteger o armazenamento em bloco na AWS)

SEC08-BP03 Automatizar a proteção de dados em repouso

Use ferramentas automatizadas para validar e impor controles de dados em repouso continuamente, por exemplo, verificar se há apenas recursos de armazenamento criptografados. Você pode automatizar a validação de que todos os volumes do EBS são criptografados com o uso do Regras do AWS Config. AWS Security Hub também pode verificar vários controles diferentes por meio de

verificações automatizadas em relação a padrões de segurança. Além disso, o Regras do AWS Config pode corrigir recursos não compatíveis automaticamente.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

Dados em repouso representam todos os dados mantidos no armazenamento não volátil por qualquer período na carga de trabalho. Isso inclui armazenamento em bloco, armazenamento de objetos, bancos de dados, arquivos, dispositivos IoT e qualquer outro meio de armazenamento no qual os dados persistam. Proteger seus dados em repouso reduz o risco de acesso não autorizado quando a criptografia e os controles de acesso adequados são implementados.

Garantir a criptografia em repouso: garanta que a única maneira de armazenar dados seja usando a criptografia. O AWS KMS se integra perfeitamente a muitos serviços da AWS para facilitar a criptografia de todos os seus dados em repouso. Por exemplo, no Amazon Simple Storage Service (Amazon S3), você pode definir a criptografia padrão em um bucket para que todos os novos objetos sejam criptografados automaticamente. Além disso, o Amazon EC2 e Amazon S3 oferecem suporte à imposição de criptografia ao definir a criptografia padrão. Você pode usar o AWS Managed Config Rules para verificar automaticamente se você está usando criptografia, por exemplo, para Volumes do EBS, instâncias do Amazon Relational Database Service (Amazon RDS)e aos Amazon S3.

Recursos

Documentos relacionados:

- Ferramentas de criptografia da AWS
- SDK de criptografia da AWS

Vídeos relacionados:

- How Encryption Works in AWS (Como a criptografia funciona na AWS)
- Securing Your Block Storage on AWS (Como proteger o armazenamento em bloco na AWS)

SEC08-BP04 Impor o 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.

Diferentes controles, incluindo acesso (usando privilégios mínimos), backups (consulte o whitepaper Confiabilidade), isolamento e versionamento, podem ajudar a proteger os dados em repouso. Deve ser feita a auditoria de acesso aos seus dados com os mecanismos de detecção abordados anteriormente neste documento, incluindo o CloudTrail e o log de nível de serviço, como os logs de acesso do Amazon Simple Storage Service (Amazon S3). Você deve inventariar quais dados são acessíveis publicamente e planejar como reduzir a quantidade de dados disponíveis ao longo do tempo. O Amazon S3 Glacier Vault Lock e o Amazon S3 Object Lock são recursos que fornecem controle de acesso obrigatório. Assim que uma política de cofre é bloqueada com a opção de conformidade, nem mesmo o usuário raiz pode alterá-la até que o bloqueio expire. O mecanismo atende aos requisitos de Books and Records Management da SEC, CFTC e FINRA. Para obter mais detalhes, consulte este whitepaper.

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

Orientação de implementação

- Aplique o controle de acesso: aplique o controle de acesso com privilégios mínimos, incluindo acesso a chaves de criptografia.
  - Introdução ao gerenciamento de permissões de acesso aos seus recursos do Amazon S3
- Dados separados com base em diferentes níveis de classificação: use diferentes de Contas da AWS para níveis de classificação de dados gerenciados pelo AWS Organizations.
  - AWS Organizations
- Analise as políticas do AWS KMS: analise o nível de acesso concedido nas políticas do AWS KMS.
  - Visão geral do gerenciamento de acesso dos recursos do AWS KMS
- Revise as permissões de objeto e de bucket do Amazon S3: revise regularmente o nível de acesso concedido nas políticas de bucket do Amazon S3. Uma das melhores práticas é não ter buckets que possam ser lidos ou gravados publicamente. Considere o uso do AWS Config para detectar buckets que estão disponíveis publicamente e do Amazon CloudFront para fornecer conteúdo do Amazon S3.
  - Regras do AWS Config
  - Amazon S3 + Amazon CloudFront: uma combinação perfeita
- Habilite o versionamento e o bloqueio de objetos do Amazon S3.
  - Usar versionamento
  - Como bloquear objetos usando o Bloqueio de objetos do Amazon S3

- Use o Amazon S3 Inventory: o Amazon S3 Inventory é uma das ferramentas que você pode usar para auditar e gerar relatórios sobre o status de replicação e criptografia de seus objetos.
  - Amazon S3 Inventory
- Revise as permissões de compartilhamento do Amazon EBS e do AMI: as permissões de compartilhamento podem autorizar que imagens e volumes sejam compartilhados com Contas da AWS externas à sua workload.
  - Como compartilhar um snapshot do Amazon EBS
  - AMIs compartilhadas

Recursos

Documentos relacionados:

- Whitepaper de detalhes criptográficos do AWS KMS

Vídeos relacionados:

- Securing Your Block Storage on AWS (Como proteger o armazenamento em bloco na AWS)

SEC08-BP05 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, use um fluxo de trabalho de gerenciamento de alterações para gerenciar instâncias do Amazon Elastic Compute Cloud (Amazon EC2) usando ferramentas em vez de permitir acesso direto ou um host traga a sua própria licença. Isso pode ser obtido usando o AWS Systems Manager Automation, que usa documentos de automação que contêm etapas que você usa para realizar tarefas. Esses documentos podem ser armazenados no controle de origem, analisados por pares antes da execução e testados detalhadamente para minimizar os riscos em comparação com o acesso ao shell. Os usuários empresariais podem ter 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.

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

Orientação de implementação

- Implemente mecanismos para manter as pessoas longe dos dados: os mecanismos incluem o uso de painéis, como o Amazon QuickSight, para exibir dados aos usuários em vez de consultar diretamente.
  - Amazon QuickSight
- Automatize o gerenciamento de configuração: execute ações remotas, aplique e valide configurações seguras automaticamente usando uma ferramenta ou um serviço de gerenciamento de configuração. Evite usar hosts traga a sua própria licença ou acessar diretamente instâncias do EC2.
  - AWS Systems Manager
  - AWS CloudFormation
  - Pipeline de CI/CD do AWS CloudFormation para modelos na AWS

Recursos

Documentos relacionados:

- Whitepaper de detalhes criptográficos do AWS KMS

Vídeos relacionados:

- How Encryption Works in AWS (Como a criptografia funciona no AWS Backup)
- Securing Your Block Storage on AWS (Como proteger o armazenamento em bloco na AWS)

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.

Práticas recomendadas

- SEC09-BP01 Implementar o gerenciamento seguro de chaves e certificados
- SEC09-BP02 Aplique a criptografia em trânsito
- SEC09-BP03 Automatizar a detecção de acesso não intencional a dados
- SEC09-BP04 Autenticar as comunicações de rede

SEC09-BP01 Implementar o gerenciamento seguro de chaves e certificados

armazene chaves de criptografia e certificados com segurança e alterne-os em intervalos de tempo apropriados com controle de acesso rigoroso. A melhor maneira de fazer isso é usar um serviço gerenciado, como o AWS Certificate Manager (ACM). Ele facilita o provisionamento, o gerenciamento e a implantação de certificados de Transport Layer Security (TLS) públicos e privados para uso com os serviços da AWS e seus recursos internos conectados. Certificados TLS são usados para proteger as comunicações de rede e estabelecer a identidade de sites pela Internet, bem como de recursos em redes privadas. O ACM se integra a recursos da AWS, como Elastic Load Balancers (ELBs), distribuições da AWS e APIs no API Gateway, além de lidar com renovações automáticas de certificados. Se você usar o ACM para implantar uma CA raiz privada, ele poderá fornecer os certificados e as chaves privadas para uso em instâncias do Amazon Elastic Compute Cloud (Amazon EC2), contêineres e outros.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Implementar o gerenciamento seguro de chaves e certificados: implemente a solução definida de gerenciamento seguro de chaves e certificados.
  - AWS Certificate Manager
  - Como hospedar e gerenciar toda uma infraestrutura de certificados privados na AWS
- Implementar protocolos seguros: use protocolos seguros que ofereçam autenticação e confidencialidade, como Transport Layer Security (TLS) ou IPsec, para reduzir o risco de violação ou perda de dados. Verifique a documentação da AWS quanto aos protocolos e segurança relevantes para os serviços que você está usando.

Recursos

Documentos relacionados:

- Documentação da AWS

SEC09-BP02 Aplique a criptografia em trânsito

Imponha o uso dos requisitos de criptografia definidos com base em padrões e recomendações apropriados para conseguir cumprir os requisitos organizacionais, legais e de conformidade.
Os serviços da AWS fornecem endpoints HTTPS usando TLS para comunicação, fornecendo

criptografia em trânsito ao se comunicar com as APIs da AWS. Protocolos não seguros, como HTTP, podem ser auditados e bloqueados em uma VPC por meio do uso de grupos de segurança. As solicitações HTTP também podem ser redirecionadas automaticamente para HTTPS no Amazon CloudFront ou em um Application Load Balancer. Você tem controle total sobre seus recursos de computação para implementar a criptografia em trânsito em seus serviços. Além disso, você pode usar a conectividade VPN em sua VPC a partir de uma rede externa para facilitar a criptografia do tráfego. Soluções de terceiros estão disponíveis no AWS Marketplace, caso você tenha requisitos especiais.

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

Orientação de implementação

- Aplique a criptografia em trânsito: os requisitos de criptografia definidos devem se basear nos mais recentes padrões e práticas recomendadas e permitir apenas protocolos seguros. Por exemplo, configure apenas um grupo de segurança para permitir o protocolo HTTPS a um Application Load Balancer ou instância do Amazon Elastic Compute Cloud (Amazon EC2).
- Configure protocolos seguros em serviços de borda: configure o HTTPS com o Amazon CloudFront e as cifras necessárias.
  - Como usar o HTTPS com o CloudFront
- Use uma VPN para conectividade externa: considere usar uma rede privada virtual (VPN) IPsec para proteger conexões ponto a ponto ou rede a rede para fornecer privacidade e integridade dos dados.
  - Conexões VPN
- Configure protocolos seguros em balanceadores de carga: habilite o ouvinte de HTTPS para proteger conexões com balanceadores de carga.
  - Ouvintes de HTTPS para o seu Application Load Balancer
- Configure protocolos seguros para instâncias: considere configurar a criptografia HTTPS em instâncias.
  - Tutorial: Configurar o servidor da Web Apache no Amazon Linux 2 para usar SSL/TLS
- Configure protocolos seguros no Amazon Relational Database Service (Amazon RDS): use Secure Socket Layer (SSL) ou Transport Layer Security (TLS) para criptografar a conexão com instâncias de banco de dados.
  - Uso de SSL para criptografar uma conexão com uma Instância de banco de dados
- Configure protocolos seguros no Amazon Redshift: configure o cluster para exigir uma conexão Secure Socket Layer (SSL) ou Transport Layer Security (TLS).

  - Configure opções de segurança para as conexões
- Configurar protocolos seguros em serviços adicionais da AWS. Para os serviços da AWS que você usa, determine os recursos de criptografia em trânsito.

Recursos

Documentos relacionados:

- Documentação da AWS

SEC09-BP03 Automatizar a detecção de acesso não intencional a dados

Use ferramentas como o Amazon GuardDuty para detectar automaticamente atividades suspeitas ou tentativas de mover dados para fora dos limites definidos. Por exemplo, o GuardDuty pode detectar atividade de leitura do Amazon Simple Storage Service (Amazon S3) que é incomum com a descoberta Exfiltration:S3/AnomalousBehavior. Além do GuardDuty, Logs de fluxo da Amazon VPC, que capturam informações de tráfego de rede, podem ser usados com o Amazon EventBridge para acionar a detecção de conexões anormais, bem-sucedidas e recusadas. Amazon S3 Access Analyzer pode ajudar a avaliar quais dados podem ser acessados por quem nos buckets do Amazon S3.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Automatizar a detecção de acesso não intencional a dados: use uma ferramenta ou um mecanismo de identificação para detectar automaticamente tentativas de mover dados fora dos limites definidos; por exemplo, para descobrir um sistema de banco de dados que esteja copiando dados para um host desconhecido.
  - Logs de fluxo da VPC
- Considerar o Amazon Macie: o Amazon Macie é um serviço de privacidade e segurança de dados totalmente gerenciado que usa machine learning e correspondência de padrões para descobrir e proteger seus dados sigilosos na AWS.
  - Amazon Macie

Recursos

Documentos relacionados:

- Logs de fluxo da VPC
- Amazon Macie

SEC09-BP04 Autenticar as comunicações de rede

Verifique a identidade das comunicações usando protocolos que oferecem suporte à autenticação, como Transport Layer Security (TLS) ou IPsec.

O uso de protocolos de rede que oferecem suporte à autenticação permite que a confiança seja estabelecida entre as partes. Isso é adicionado à criptografia usada no protocolo para reduzir o risco de as comunicações serem alteradas ou interceptadas. Protocolos comuns que implementam a autenticação incluem Transport Layer Security (TLS), que é usado em muitos serviços da AWS, e o IPsec, que é usado na AWS Virtual Private Network (AWS VPN).

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

Orientação de implementação

- Implemente protocolos seguros: use protocolos seguros que ofereçam autenticação e confidencialidade, como TLS ou IPsec, para reduzir o risco de violação ou perda de dados. Consulte a Documentação da AWS quanto aos protocolos e à segurança relevantes para os serviços que você está usando.

Recursos

Documentos relacionados:

- Documentação da AWS

Resposta a incidentes

Pergunta

- SEC 10 Como você prevê, responde e se recupera de 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.

Práticas recomendadas

- SEC10-BP01 Identificar o pessoal-chave e os recursos externos
- SEC10-BP02 Desenvolver planos de gerenciamento de incidentes
- SEC10-BP03 Preparar recursos forenses
- SEC10-BP04 Automatizar a capacidade de contenção
- SEC10-BP05 Acesso pré-provisionado
- SEC10-BP06 Ferramentas pré-implantação
- SEC10-BP07 Promover dias de jogo

SEC10-BP01 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.

Para definir sua abordagem de resposta a incidentes na nuvem, com a participação de outras equipes (como consultoria jurídica, liderança, partes interessadas de negócios, serviços do AWS Support e outras), você deve identificar as principais partes interessadas, pessoal e contatos relevantes. Para reduzir a dependência e diminuir o tempo de resposta, certifique-se de que sua equipe, equipes de segurança especializadas e respondentes sejam instruídos sobre os serviços que você usa e tenham a oportunidade de praticar.

É recomendável identificar parceiros externos de segurança da AWS que possam fornecer experiência externa e uma perspectiva diferente para aumentar seus recursos de resposta. Os parceiros de segurança confiáveis podem ajudá-lo a identificar possíveis riscos ou ameaças com os quais você talvez não esteja familiarizado.

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

Orientação de implementação

- Identifique o pessoal-chave em sua organização: mantenha uma lista de contatos da sua organização que você precisaria acionar para responder e recuperar-se de um incidente.

- Identifique parceiros externos: entre em contato com parceiros externos, se necessário, que possam ajudar você a responder e se recuperar de um incidente.

Recursos

Documentos relacionados:

- AWS Incident Response Guide (Guia de resposta a incidentes da AWS)

Vídeos relacionados:

- Prepare for and respond to security incidents in your AWS environment (Prepare-se e responda a incidentes de segurança no ambiente da AWS)

Exemplos relacionados:

SEC10-BP02 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.

Nível de risco exposto se essa prática recomendada não for estabelecida: alto

Orientação para implementação

Um plano de gerenciamento de incidentes é fundamental para responder, mitigar e se recuperar de possíveis impactos de incidentes de segurança. Um plano de gerenciamento de incidentes é um processo estruturado de identificação, correção e resposta em tempo hábil a incidentes de segurança.

A nuvem tem muitos dos mesmos requisitos e perfis operacionais encontrados em um ambiente on-premises. Ao criar um plano de gerenciamento de incidentes, é importante definir estratégias de resposta e recuperação que se alinhem melhor aos seus resultados empresariais e requisitos de conformidade. Por exemplo, se você opera workloads na AWS em conformidade com o FedRAMP nos Estados Unidos, é útil aderir ao Guia de tratamento de segurança de computadores NIST SP 800-61. Da mesma forma, ao operar workloads com dados europeus de PII (informações de identificação pessoal), considere cenários como a forma como você deve se proteger e responder

a incidentes relacionados à residência de dados, conforme exigido pela Regulamentação Geral de Proteção de Dados (GDPR) da UE.

Ao criar um plano de gerenciamento de incidentes para suas workloads em operação na AWS, comece com o Modelo de responsabilidade compartilhada da AWS, para elaborar uma abordagem de defesa profunda em relação à resposta a incidentes. Nesse modelo, a AWS gerencia a segurança da nuvem, e você é responsável pela segurança na nuvem. Isso significa que você mantém o controle e é responsável pelos controles de segurança que escolhe implementar. O AWS Security Incident Response Guide (Guia de resposta a incidentes de segurança da AWS) detalha os conceitos e as orientações básicas para criar um plano de gerenciamento de incidentes centrado na nuvem.

Um plano de gerenciamento de incidentes eficaz deve ser continuamente iterado e permanecer atualizado com relação às suas metas de operações de nuvem. Considere o uso dos planos de implementação detalhados abaixo, à medida que cria e evolui seu plano de gerenciamento de incidentes.

- Promova instrução e treinamento para a resposta a incidentes: quando ocorre um desvio de sua referência básica definida (por exemplo, um erro de implantação ou de configuração), você pode precisar investigar e dar uma resposta. Para fazer isso com sucesso, é necessário entender quais controles e recursos podem ser usados para a resposta ao incidente de segurança em seu ambiente da AWS, bem como os processos que você deve considerar para preparar, instruir e treinar suas equipes de nuvem que participam da resposta a um incidente.
  - Manuais e runbooks são mecanismos eficazes para criar consistência no treinamento de como responder a incidentes. Comece criando uma lista inicial de procedimentos executados com frequência durante a resposta a um incidente e continue a iterar à medida que você aprende ou usa novos procedimentos.
  - Socialize os manuais e runbooks por meio de dias de jogos agendados. Durante os dias de jogos, simule a resposta a incidentes em um ambiente controlado para que sua equipe possa se lembrar de como responder e para verificar se as equipes envolvidas na resposta a incidentes conhecem bem os fluxos de trabalho. Revise os resultados do evento simulado para identificar melhorias e determinar a necessidade de mais treinamento ou ferramentas adicionais.
  - A segurança deve ser considerada um trabalho de todos. Crie um conhecimento coletivo do processo de gerenciamento de incidentes envolvendo todo o pessoal que geralmente opera suas workloads. Isso inclui todos os aspectos de sua empresa: operações, teste, desenvolvimento, segurança, operações empresariais e líderes da empresa.

- Documente o plano de gerenciamento de incidentes: Documente as ferramentas e os processos para registrar, tomar medidas, comunicar o andamento e notificar sobre os incidentes ativos. A meta do plano de gerenciamento de incidentes é verificar se a operação normal é restaurada o mais rapidamente possível, se o impacto empresarial é minimizado e se todas as partes interessadas são informadas. Exemplos de incidentes incluem (mas não se restringem a) perda ou degradação da conectividade de rede, uma API ou um processo que não responde, uma tarefa programada não realizada (por exemplo, falha na aplicação de patches), indisponibilidade de serviço ou dados da aplicação, interrupção não planejada do serviço devido a eventos de segurança, vazamento de credenciais ou erros de configuração.
  - Identifique o proprietário principal responsável pela resolução do incidente, como o proprietário da workload. Tenha orientações claras de quem vai gerenciar o incidente e de como a comunicação será tratada. Quando você tem mais de uma parte participando do processo de resolução do incidente, como um fornecedor externo, considere a criação de uma matriz de responsabilidade (RACI), detalhando as funções e responsabilidades de várias equipes ou pessoas necessárias para a resolução do incidente.

    Uma matriz de RACI detalha o seguinte:

    - R: parte responsável que faz o trabalho para concluir a tarefa.
    - A: parte atribuída com autoridade financeira pela conclusão bem-sucedida da tarefa específica.
    - C: parte consultada cujas opiniões são procuradas, geralmente como especialistas no assunto.
    - I: parte informada que é notificada sobre o andamento, geralmente apenas depois da conclusão da tarefa ou dos resultados.
- Categorize os incidentes: definir e categorizar incidentes com base em pontuações de gravidade e impacto permite uma abordagem estruturada para fazer a triagem e solucionar os incidentes. As recomendações a seguir ilustram uma matriz de urgência do impacto à resolução para quantificar um incidente. Por exemplo, um incidente de baixo impacto e baixa urgência é considerado um incidente de baixa gravidade.
  - Alto (H): sua empresa é afetada significativamente. Funções críticas de sua aplicação relacionadas aos recursos da AWS ficam indisponíveis. Classificação reservada para a maioria dos eventos críticos que afetam os sistemas de produção. O impacto do incidente aumenta rapidamente, fazendo com que a correção precise ocorrer o mais rapidamente possível.
  - Médio (M): uma aplicação ou um serviço da empresa relacionado aos recursos da AWS é afetado moderadamente e funciona em um estado degradado. Aplicações que contribuem com

os objetivos do nível de serviço (SLOs) são afetadas dentro dos limites do Acordo de Serviço (SLA). Os sistemas podem ser operados com capacidade reduzida sem muito impacto financeiro e de reputação.

- Baixo (L): funções não críticas de sua aplicação ou serviço empresarial relacionado aos recursos da AWS são afetadas. Os sistemas podem ser operados com capacidade reduzida com impacto financeiro e de reputação mínimo.

- Padronize os controles de segurança: a meta da padronização dos controles de segurança é obter consistência, rastreabilidade e repetibilidade com relação aos resultados operacionais. Promova a padronização em atividades principais que sejam críticas para a resposta a incidentes, como:
  - Gerenciamento de identidade e acesso: estabeleça mecanismos para controlar o acesso aos dados e gerenciar privilégios para identidades humanas e de máquina. Amplie o gerenciamento de sua própria identidade e acesso para a nuvem, usando segurança federada com autenticação única e privilégios baseados em funções para otimizar o gerenciamento de acesso. Para ver as práticas recomendadas e os planos de melhoria para padronizar o gerenciamento de acesso, consulte a seção de gerenciamento de identidade e acesso do whitepaper Security Pillar (Pilar de segurança).
  - Gerenciamento de vulnerabilidades: estabeleça mecanismos para identificar vulnerabilidades em seu ambiente da AWS que tenha a probabilidade de ser usado por invasores para comprometer e fazer uso indevido de seu sistema. Implemente controles de prevenção e detecção, como mecanismos de segurança, para responder e mitigar o possível impacto dos incidentes de segurança. Padronize processos como a modelagem de ameaças como parte do ciclo de vida de entrega de aplicações e compilação de infraestrutura.
  - Gerenciamento de configurações: Defina configurações padrão e automatize procedimentos para implantar recursos na Nuvem AWS. Padronizar o provisionamento de recursos e infraestrutura ajuda a mitigar o risco de erros de configuração por meio de implantações incorretas ou erros de configuração acidentais por humanos. Consulte a seção de princípios do projeto do whitepaper Operational Excellence Pillar (Pilar de excelência operacional) a fim de obter orientações e planos de melhoria para implementar esse controle.
  - Registro e monitoramento do controle de auditoria: implemente mecanismos para monitorar seus recursos em busca de falhas, degradação do desempenho e problemas de segurança. Padronizar esses controles também fornece trilhas de atividades de auditoria que ocorrem em seu sistema, ajudando a fazer a triagem e a correção dos problemas em tempo hábil. As práticas recomendadas em SEC04 (“Como você detecta e investiga eventos de segurança?”) fornecem orientações de implementação desse controle.

- Use a automação: a automação permite solucionar o incidente em larga escala e em tempo hábil. A AWS oferece vários serviços para automatização no contexto da estratégia de resposta a incidentes. Concentre-se em encontrar o equilíbrio adequado entre a automação e a intervenção manual. À medida que você cria sua resposta a incidentes em manuais e runbooks, automatize as etapas repetíveis. Use os serviços da AWS, como o AWS Systems Manager Incident Manager para solucionar incidentes de TI mais rapidamente. Use ferramentas de desenvolvedor para fornecer controle de versão e automatizar o Amazon Machine Images (AMI) e implantações de infraestrutura como código (IaC) sem intervenção humana. Quando aplicável, automatize a detecção e a avaliação de conformidade usando serviços gerenciados, como o Amazon GuardDuty, o Amazon Inspector, o AWS Security Hub, o AWS Config e o Amazon Macie. Otimize os recursos de detecção com machine learning, como o Amazon DevOps Guru, para detectar padrões de operação anormais antes que eles ocorram.
- Realize uma análise da causa raiz e coloque em prática as lições aprendidas: implemente mecanismos para guardar as lições aprendidas como parte de uma avaliação após a resposta a incidentes. Quando a causa raiz de um incidente revela um defeito maior, uma falha de projeto, um erro de configuração ou uma possibilidade de recorrência, ele é classificado como um problema. Nesses casos, analise e resolva o problema para minimizar a interrupção de operações normais.

Recursos

Documentos relacionados:

- AWS Security Incident Response Guide (Guia de resposta a incidentes de segurança da AWS)
- NIST: Guia de tratamento de incidentes de segurança de computadores

Vídeos relacionados:

- Automating Incident Response and Forensics in AWS (Automação de resposta a incidentes e investigações forenses na AWS)
- Guia DIY (faça você mesmo) para runbooks, relatórios de incidentes e resposta a incidentes
- Prepare for and respond to security incidents in your AWS environment (Prepare-se e responda a incidentes de segurança no ambiente da AWS)

Exemplos relacionados:

- Lab: Incident Response Playbook with Jupyter - AWS IAM (Laboratório: Manual de resposta a incidentes com o Jupyter: AWS IAM)
- Lab: Incident Response with AWS Console and CLI (Laboratório: resposta a incidentes com o console e a CLI da AWS)

SEC10-BP03 Preparar recursos forenses

É importante que os respondentes a incidentes entendam quando e como a investigação forense se encaixa no plano de resposta. A organização deve definir quais evidências são coletadas e quais ferramentas são usadas no processo. Identifique e prepare recursos de investigação forense adequados, incluindo especialistas externos, ferramentas e automação. Uma decisão importante que você deve tomar inicialmente é se você coletará dados de um sistema ativo. Alguns dados, como o conteúdo da memória volátil ou conexões de rede ativas, serão perdidos se o sistema for desligado ou reinicializado.

A equipe de resposta pode combinar ferramentas, como AWS Systems Manager, Amazon EventBridge e AWS Lambda, para executar automaticamente ferramentas forenses em um sistema operacional e espelhamento de tráfego de VPC para obter uma captura de pacote de rede e coletar evidências não persistentes. Conduza outras atividades, como análise de log ou análise de imagens de disco, em uma conta de segurança dedicada com estações de trabalho forenses personalizadas e ferramentas acessíveis a seus respondentes.

Envie logs relevantes rotineiramente para um armazenamento de dados que oferece alta durabilidade e integridade. Os respondentes devem ter acesso a esses logs. A AWS oferece várias ferramentas que podem facilitar a investigação de logs, como Amazon Athena, Amazon OpenSearch Service (OpenSearch Service) e Amazon CloudWatch Logs Insights. Além disso, preserve a evidência com segurança usando o Amazon Simple Storage Service (Amazon S3) Object Lock. Esse serviço segue o modelo de gravação única e várias leituras (WORM) e evita que objetos sejam excluídos ou substituídos por um período definido. Como as técnicas de investigação pericial exigem treinamento especializado, pode ser necessário envolver especialistas externos.

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

Orientação de implementação

- Identifique os recursos forenses: pesquise os recursos de investigação forense da sua organização, as ferramentas disponíveis e os especialistas externos.
- Automatização de resposta a incidentes e forense

Recursos

Documentos relacionados:

- How to automate forensic disk collection in AWS (Como automatizar a coleta de disco forense na AWS)

SEC10-BP04 Automatizar a capacidade de contenção

Automatize os recursos de contenção e recuperação de incidentes para reduzir o tempo de resposta e o impacto organizacional.

Depois de criar e praticar os processos e as ferramentas com seus playbooks, você poderá desconstruir a lógica de uma solução baseada em código, que pode ser usada como ferramenta por muitos respondentes para automatizar a resposta e remover variações ou suposições dos respondentes. Isso pode acelerar o ciclo de vida de uma resposta. O próximo objetivo é permitir a total automatização desse código por meio da invocação dos alertas ou dos eventos por si mesmo, e não por um respondente humano, para criar uma resposta orientada por eventos. Esses processos também devem adicionar automaticamente dados relevantes aos sistemas de segurança. Por exemplo, um incidente envolvendo o tráfego de um endereço IP indesejado pode preencher automaticamente uma lista de bloqueios do AWS WAF ou um grupo de regras de firewall de rede para evitar mais atividades.

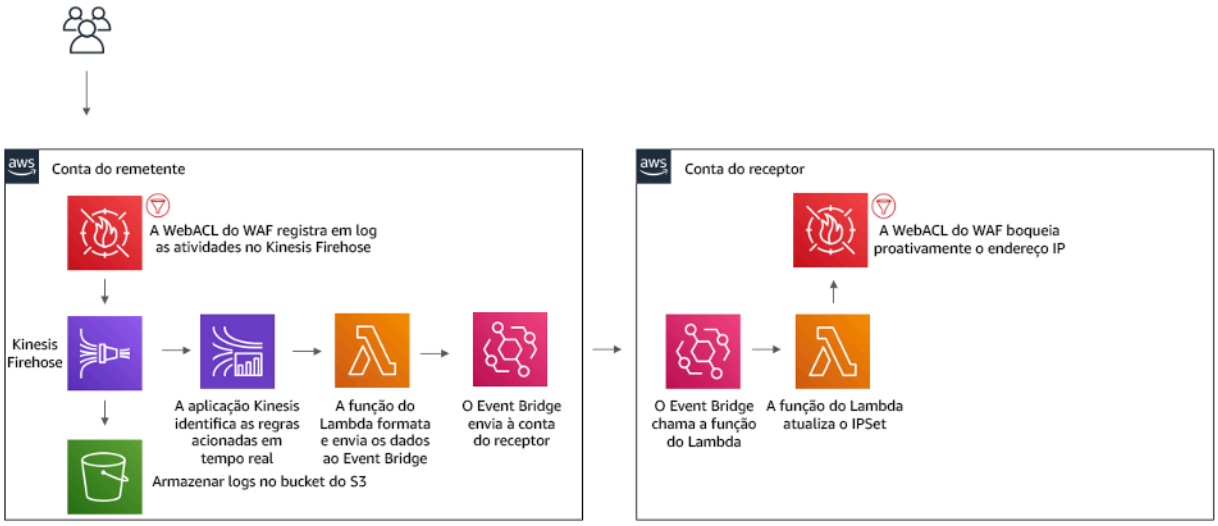


Figura 3: O AWS WAF automatiza o bloqueio de endereços IP maliciosos conhecidos.

Com um sistema de resposta orientado por eventos, um mecanismo de detecção aciona um mecanismo responsivo para corrigir automaticamente o evento. Você pode usar recursos de

resposta orientados por eventos para reduzir o tempo de retorno entre os mecanismos de detecção e os mecanismos responsivos. Para criar essa arquitetura orientada por eventos, é possível usar o AWS Lambda, que é um serviço de computação sem servidor que executa o código em resposta a eventos e gerencia automaticamente os recursos computacionais subjacentes. Por exemplo, suponha que você tenha uma conta da AWS com o serviço AWS CloudTrail habilitado. Se o AWS CloudTrail estiver desabilitado (por meio da chamada de API `cloudtrail:StopLogging`), você pode usar o Amazon EventBridge para monitorar o evento `cloudtrail:StopLogging` específico e invocar uma função do AWS Lambda para chamar `cloudtrail:StartLogging` para reiniciar o registro em log.

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

Orientação de implementação

Automatize a capacidade de contenção.

Recursos

Documentos relacionados:

- AWS Incident Response Guide (Guia de resposta a incidentes da AWS)

Vídeos relacionados:

- Prepare for and respond to security incidents in your AWS environment (Prepare-se e responda a incidentes de segurança no ambiente da AWS)

SEC10-BP05 Acesso pré-provisionado

Verifique se os respondentes a incidentes têm o acesso correto pré-provisionado na AWS para reduzir o tempo de investigação necessário até a recuperação.

Antipadrões comuns:

- Uso da conta raiz para a resposta a incidentes.
- Alteração de contas de usuário existentes.
- Manipulação de permissões do IAM diretamente ao fornecer elevação de privilégios just-in-time.

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

Orientação para implementação

A AWS recomenda reduzir ou eliminar a dependência de credenciais de longa duração sempre que possível, dando preferência a credenciais temporárias e a mecanismos de escalação de privilégios just-in-time. As credenciais de longa duração são propensas a riscos de segurança e aumentam a sobrecarga operacional. Para a maioria das tarefas de gerenciamento, bem como tarefas de resposta a incidentes, recomendamos a implementação da federação de identidades junto com a escalação temporária para acesso administrativo. Nesse modelo, um usuário solicita elevação a um nível superior de privilégio (como um perfil de resposta a incidentes) e, considerando que ele seja elegível para a elevação, a solicitação é enviada a um aprovador. Se a solicitação for aprovada, o usuário receberá um conjunto de credenciais temporárias da AWS, que podem ser usadas para concluir suas tarefas. Depois que essas credenciais expirarem, o usuário deve enviar uma nova solicitação de elevação.

Recomendamos o uso da escalação de privilégio temporária para a maioria dos cenários de resposta a incidentes. A maneira correta de fazer isso é com o uso do AWS Security Token Service e de políticas de sessão para definir o escopo de acesso.

Há cenários em que as identidades federadas não estão disponíveis, como:

- Interrupção relacionada a um provedor de identidades (IdP) comprometido.
- Erro de configuração ou erro humano causando uma falha no sistema de gerenciamento de acesso federado.
- Atividade mal-intencionada, como um evento de negação de serviço distribuído (DDoS) ou indisponibilidade de renderização do sistema.

Nos casos anteriores, deverá haver um acesso de emergência de breaking-glass configurado para permitir a investigação e a correção em tempo hábil dos incidentes. Recomendamos a utilização de um usuário do IAM com as permissões apropriadas para realizar tarefas e acessar os recursos da AWS. Use as credenciais raiz somente para tarefas que exijam o acesso do usuário raiz. Para verificar se os respondentes de um incidente têm o nível de acesso correto à AWS e a outros sistemas relevantes, recomendamos o pré-provisionamento de contas de usuário dedicadas. As contas de usuário exigem acesso privilegiado e devem ser estritamente controladas e monitoradas. As contas devem ser criadas com os menores privilégios exigidos para realizar as tarefas necessárias, e o nível de acesso deve ser baseado nos manuais criados como parte do plano de gerenciamento de incidentes.

Utilize perfis e usuários dedicados e com propósito específico como uma prática recomendada. Escalar temporariamente o acesso de usuários ou perfis por meio da adição de políticas do IAM não deixa claro qual é o acesso que os usuários tinham durante o incidente, e há um risco de que os privilégios escalados não sejam revogados.

É importante remover o máximo de dependências possível para verificar se o acesso pode ser obtido com o maior número possível de cenários de falha. Para apoiar isso, crie um manual para verificar se os usuários de resposta a incidentes são criados como usuários do AWS Identity and Access Management em uma conta de segurança dedicada, e não são gerenciados por nenhuma solução de autenticação única (SSO) ou federação. Cada respondente individual deve ter sua própria conta nomeada. A configuração da conta deve aplicar uma política de senha forte e a autenticação multifator (MFA). Se os manuais de resposta a incidentes só exigem acesso ao AWS Management Console, o usuário não deve ter chaves de acesso configuradas e deve ser proibido explicitamente de criar chaves de acesso. Isso pode ser configurado com políticas do IAM ou políticas de controle de serviços (SCPs), conforme mencionado nas Práticas recomendadas de segurança da AWS para SCPs do AWS Organizations. Os usuários não devem ter privilégios além da capacidade de assumir perfis de resposta a incidentes em outras contas.

Durante um incidente, pode ser necessário conceder acesso a outros indivíduos internos ou externos para apoiar a investigação, a correção ou as atividades de recuperação. Nesse caso, use o mecanismo do manual mencionado anteriormente, e deve haver um processo para verificar se qualquer acesso adicional foi revogado imediatamente após a conclusão do incidente.

Para verificar se o uso de perfis de resposta a incidentes pode ser monitorado e auditado corretamente, é essencial que as contas de usuário do IAM criadas para esse fim não sejam compartilhadas entre indivíduos e que o usuário raiz da Conta da AWS não seja utilizado, a menos que isso seja exigido para uma tarefa específica. Se o usuário raiz for exigido (por exemplo, quando o acesso do IAM a uma conta específica estiver indisponível), use um processo distinto com um manual disponível para verificar a disponibilidade da senha e do token de MFA do usuário raiz.

Para configurar as políticas do IAM para os perfis de resposta a incidentes, considere o uso do IAM Access Analyzer para gerar políticas baseadas em logs do AWS CloudTrail. Para fazer isso, conceda acesso de administrador ao perfil de resposta a incidentes em uma conta de não produção e execute de acordo com os manuais. Depois da conclusão, pode ser criada uma política que permita somente as ações realizadas. Essa política pode ser então aplicada a todos os perfis de resposta a incidentes em todas as contas. Você pode criar uma política do IAM separada para cada manual a fim de facilitar o gerenciamento e a auditoria. Exemplos de manuais podem incluir planos de resposta para ransomware, violações de dados, perda de acesso da produção, dentre outros cenários.

Use as contas de usuário de resposta a incidentes para assumir funções do IAM de resposta a incidentes em outras Contas da AWS. Esses perfis também devem ser configurados para só poderem ser assumidos por usuários na conta de segurança, e o relacionamento de confiança deve exigir que a entidade principal que está fazendo a chamada seja autenticada com MFA. Os perfis devem usar políticas do IAM com escopo estritamente definido para controlar o acesso. Garanta que todas as solicitações `AssumeRole` para esses perfis estejam conectadas no CloudTrail e sejam alertadas, e que as ações realizadas usando esses perfis sejam registradas.

É altamente recomendável que as contas de usuário do IAM e os perfis do IAM sejam claramente nomeados para permitir que sejam encontrados com facilidade nos logs do CloudTrail. Um exemplo disso seria nomear as contas do IAM como *<USER_ID>*`-BREAK-GLASS` e os perfis do IAM como `BREAK-GLASS-ROLE`.

O CloudTrail é usado para registrar as atividades da API em suas contas da AWS e deve ser usado para configurar alertas sobre o uso dos perfis de resposta a incidentes. Consulte a publicação do blog sobre como configurar alertas quando as chaves raiz são usadas. As instruções podem ser modificadas para configurar a métrica do Amazon CloudWatch filtro a filtro em eventos `AssumeRole` relacionados ao perfil do IAM de resposta a incidentes:

```
{ $.eventName = "AssumeRole" && $.requestParameters.roleArn =
 "<INCIDENT_RESPONSE_ROLE_ARN>" && $.userIdentity.invokedBy NOT EXISTS && $.eventType !
= "AwsServiceEvent" }
```

Como é provável que os perfis de resposta a incidentes tenham um alto nível de acesso, é importante que esses alertas sejam transmitidos a um grande grupo e que sejam tomadas atitudes rapidamente.

Durante um incidente, é possível que um respondente possa exigir acesso a sistemas que não são protegidos diretamente pelo IAM. Isso pode incluir instâncias do Amazon Elastic Compute Cloud, bancos de dados do Amazon Relational Database Service ou plataformas de software como serviço (SaaS). É altamente recomendável que, em vez de usar protocolos nativos, como SSH ou RDP, o AWS Systems Manager Session Manager seja usado para todo acesso administrativo a instâncias do Amazon EC2. Esse acesso pode ser controlado usando o IAM, que é protegido e auditado. Também pode ser possível automatizar partes de seus manuais usando os documentos do AWS Systems Manager Run Command, o que pode reduzir os erros do usuário e melhorar o tempo de recuperação. Para acesso aos bancos de dados e a ferramentas de terceiros, recomendamos armazenar as credenciais de acesso no AWS Secrets Manager e conceder acesso aos perfis de respondente a incidentes.

Por fim, o gerenciamento das contas de usuário do IAM de resposta a incidentes deve ser adicionado aos seus processos de junção, migração e saída, além de ser revisado e testado periodicamente visando confirmar se somente o acesso pretendido é permitido.

Recursos

Documentos relacionados:

- Managing temporary elevated access to your AWS environment (Gerenciamento de acesso elevado temporário ao seu ambiente da AWS)
- AWS Security Incident Response Guide (Guia de resposta a incidentes de segurança da AWS)
- AWS Elastic Disaster Recovery
- AWS Systems Manager Incident Manager
- Setting an account password policy for IAM users (Definição de uma política de senhas de contas para usuários do IAM)
- Using multi-factor authentication (MFA) in AWS (Uso da autenticação multifator (MFA) na AWS)
- Configuring Cross-Account Access with MFA (Configuração do acesso entre contas com MFA)
- Using IAM Access Analyzer to generate IAM policies (Uso do IAM Access Analyzer para gerar políticas do IAM)
- Best Practices for AWS Organizations Service Control Policies in a Multi-Account Environment (Práticas recomendadas para políticas de controle de serviço do AWS Organizations em um ambiente de várias contas)
- How to Receive Notifications When Your AWS Account's Root Access Keys Are Used (Como receber notificações quando as chaves de acesso raiz da sua conta da AWS são usadas)
- Create fine-grained session permissions using IAM managed policies (Criar permissões de sessão refinadas usando políticas gerenciadas pelo IAM)

Vídeos relacionados:

- Automating Incident Response and Forensics in AWS (Automação de resposta a incidentes e investigações forenses na AWS)
- Guia DIY (faça você mesmo) para runbooks, relatórios de incidentes e resposta a incidentes
- Prepare for and respond to security incidents in your AWS environment (Prepare-se e responda a incidentes de segurança no ambiente da AWS)

Exemplos relacionados:

- Lab: AWS Account Setup and Root User (Laboratório: usuário raiz e configuração de conta da AWS)
- Lab: Incident Response with AWS Console and CLI (Laboratório: resposta a incidentes com o console e a CLI da AWS)

SEC10-BP06 Ferramentas pré-implantação

garanta que o pessoal de segurança tenha as ferramentas certas pré-implantadas na AWS para reduzir o tempo de investigação até a recuperação.

Para automatizar as funções de engenharia e operações de segurança, você pode usar um conjunto abrangente de APIs e ferramentas da AWS. Você pode automatizar totalmente os recursos de gerenciamento de identidade, segurança de rede, proteção de dados e monitoramento e disponibilizá-los com métodos populares de desenvolvimento de software já em vigor. Quando você cria a automação da segurança, seu sistema pode monitorar, analisar e iniciar uma resposta, em vez de fazer com que as pessoas monitorem a sua posição de segurança e reajam manualmente a eventos. Uma maneira eficaz de fornecer automaticamente dados de log relevantes e pesquisáveis em todos os serviços da AWS para seus atendentes de incidentes é habilitar o Amazon Detective.

Se as equipes de resposta a incidentes continuarem a responder aos alertas da mesma forma, há o risco de se acostumarem aos alertas. Com o passar do tempo, a equipe pode se tornar dessensibilizada para alertas e cometer erros ao lidar com situações comuns ou perder alertas incomuns. A automação ajuda a evitar a exaustão de alertas usando funções que processam alertas repetitivos e comuns, permitindo que as pessoas lidem com incidentes confidenciais e exclusivos. A integração de sistemas de detecção de anomalias, como Amazon GuardDuty, AWS CloudTrail Insights e Amazon CloudWatch Anomaly Detection, pode reduzir a carga de alertas baseados em limites comuns.

Você pode melhorar os processos manuais com a automatização programática das etapas do processo. Depois de definir o padrão de correção para um evento, você pode decompor esse padrão em lógica acionável e desenvolver o código para executar essa lógica. Os respondentes podem executar esse código para corrigir o problema. Com o passar do tempo, você pode automatizar mais e mais etapas e, por fim, lidar automaticamente com classes inteiras de incidentes comuns.

As ferramentas executadas no sistema operacional da instância do Amazon Elastic Compute Cloud (Amazon EC2) devem ser avaliadas com Run Command do AWS Systems Manager, que permite administrar instâncias de forma remota e segura usando um agente que você instala no sistema

operacional de instância do Amazon EC2. Ele requer o Systems Manager Agent (SSM Agent), que é instalado por padrão em muitas imagens de máquina da Amazon (AMIs). Porém, lembre-se de que, se uma instância for comprometida, nenhuma resposta das ferramentas ou dos agentes que ela executa será considerada confiável.

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

Orientação de implementação

- Pré-implante as ferramentas: para que uma resposta apropriada possa ser dada a um incidente, assegure que a equipe de segurança tenha as ferramentas certas pré-implantadas na AWS.
  - Laboratório: Resposta a incidentes com a CLI e o AWS Management Console
  - Playbook de resposta a incidentes com o Jupyter: AWS IAM
  - AWS Security Automation
- Implemente a marcação de recursos: marque recursos com informações, como um código para o recurso sob investigação, para identificar recursos durante um incidente.
  - Estratégias de marcação da AWS

Recursos

Documentos relacionados:

- AWS Incident Response Guide (Guia de resposta a incidentes da AWS)

Vídeos relacionados:

- DIY guide to runbooks, incident reports, and incident response

SEC10-BP07 Promover dias de jogo

dias de jogos, também conhecidos como simulações ou exercícios, são eventos internos que oferecem uma oportunidade estruturada para praticar seus planos e procedimentos de gerenciamento de incidentes em um cenário realista. Esses eventos devem treinar os respondentes usando as mesmas ferramentas e técnicas que seriam usadas em um cenário real, inclusive imitando ambientes reais. Os dias de jogos abrangem fundamentalmente a preparação e a melhoria iterativa dos recursos de resposta. Alguns dos motivos pelos quais você pode encontrar valor na execução de atividades do dia do jogo incluem:

- Validar a prontidão
- Desenvolver confiança - aprendizado com simulações e equipes de treinamento
- Seguir a conformidade ou obrigações contratuais
- Gerar artefatos para credenciamento
- Ser ágil - melhorias incrementais
- Tornar-se mais rápido e melhorar ferramentas
- Refinar a comunicação e a escalação
- Ter tranquilidade diante de eventos raros e inesperados

Por esses motivos, o valor derivado da participação em uma atividade de simulação aumenta a eficácia da organização durante eventos estressantes. Desenvolver uma atividade de simulação que seja realista e benéfica pode ser um exercício difícil. Embora testar seus procedimentos ou automação para eventos bem compreendidos tenha certas vantagens, é igualmente valioso participar de atividades criativas de Simulações de resposta a incidentes de segurança (SIRS) para preparar você para o inesperado e melhorar continuamente.

Crie simulações personalizadas sob medida para ambientes, equipes e ferramentas. Encontre um problema e crie a simulação com base nele. Pode ser algo como uma credencial vazada, um servidor se comunicando com sistemas indesejados ou uma configuração incorreta que resulta em exposição não autorizada. Identifique engenheiros que estão familiarizados com a organização para criar o cenário e outro grupo para participar. O cenário deve ser realista e desafiador o suficiente para ser relevante. Ele deve incluir a oportunidade de colocar na prática registros em log, notificações, escalonamentos e execução de runbooks ou automação. Durante a simulação, os respondentes devem exercitar suas habilidades técnicas e organizacionais, e os líderes devem participar para desenvolver suas habilidades de gerenciamento de incidentes. No final da simulação, comemore os esforços da equipe e procure formas de iterar, repetir e expandir para outras simulações.

A AWS criou modelos de runbook de resposta a incidentes que você pode usar não apenas para preparar os esforços de resposta, mas também como base para uma simulação. Ao planejar, uma simulação pode ser dividida em cinco fases.

Coleta de provas: nesta fase, uma equipe receberá alertas por diversos meios, como sistema interno de tíquetes, alertas de ferramentas de monitoramento, dicas anônimas ou até notícias públicas. Em seguida, as equipes começam a revisar os logs de infraestrutura e aplicações para determinar a

origem do comprometimento. Esta etapa também deve envolver escalações internas e liderança de incidentes. Após a identificação, as equipes passam a conter o incidente.

Contenção do incidente: as equipes terão determinado que houve um incidente e estabelecido a fonte do comprometimento. As equipes agora devem tomar medidas para contê-lo, por exemplo, desabilitando credenciais comprometidas, isolando um recurso de computação ou revogando a permissão de uma função.

Erradicação do incidente: agora que o incidente foi contido, as equipes trabalharão para mitigar quaisquer vulnerabilidades em aplicações ou configurações de infraestrutura que eram suscetíveis ao comprometimento. Isso pode incluir a alternância de todas as credenciais usadas para uma workload, a modificação de listas de controle de acesso (ACLs) ou a alteração das configurações de rede.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Executar dias de jogo: execute eventos simulados de resposta a incidentes (dias de jogo) para diferentes ameaças que envolvem equipe e gerenciamento importantes.
- Guardar as lições aprendidas: lições aprendidas durante os dias de jogo devem fazer parte de uma análise de feedback para melhorar seus processos.

Recursos

Documentos relacionados:

- AWS Incident Response Guide (Guia de resposta a incidentes da AWS)
- AWS Elastic Disaster Recovery

Vídeos relacionados:

- DIY guide to runbooks, incident reports, and incident response

Confiabilidade

Tópicos

- Fundamentos

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

Fundamentos

Perguntas

- REL 1 Como você gerencia as cotas e restrições de serviço?
- REL 2 Como você planeja sua topologia de rede?

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.

Práticas recomendadas

- REL01-BP01 Conhecimento das cotas e restrições de serviço
- REL01-BP02 Gerenciar cotas de serviço de várias contas e regiões
- REL01-BP03 Acomodar as cotas e as restrições fixas de serviço por meio da arquitetura
- REL01-BP04 Monitorar e gerenciar cotas
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL01-BP06 Certificar-se de que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover

REL01-BP01 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.

O Service Quotas é um serviço da AWS que ajuda você a gerenciar as cotas de mais de 100 serviços da AWS em um único local. Além de pesquisar os valores de cotas, você também pode

solicitar e acompanhar aumentos de cota no console do Service Quotas ou por meio do AWS SKD. O AWS Trusted Advisor oferece uma verificação de cotas de serviço que exibe o uso e as cotas para certos aspectos de alguns serviços. As cotas de serviço padrão por serviço também estão na documentação da AWS com base no respectivo serviço. Por exemplo, consulte Cotas da Amazon VPC. Os limites de taxa para APIs limitadas são definidos no próprio API Gateway por meio da configuração de um plano de uso. Outros limites definidos como configuração em seus respectivos serviços incluem IOPS provisionadas, armazenamento do RDS alocado e alocações de volume do EBS. O Amazon Elastic Compute Cloud (Amazon EC2) tem seu próprio painel de limites de serviço, que pode ajudar você a gerenciar sua instância, o Amazon Elastic Block Store (Amazon EBS) e os limites de endereços IP elásticos. Se você tiver um caso de uso em que as cotas de serviço afetam a performance do seu aplicativo e elas não forem ajustadas às suas necessidades, entre em contato com o AWS Support para ver se há mitigações.

Antipadrões comuns:

- Implantar uma workload sem levar em consideração as cotas de serviço referentes aos serviços da AWS usados.
- Projetar uma workload sem investigar e acomodar as restrições de design dos serviços da AWS.
- Implantar uma workload com uso significativo que substitui uma workload existente conhecida sem antes configurar as cotas necessárias ou entrar em contato com o AWS Support.
- Planejar um evento para direcionar o tráfego para sua workload, mas não configurar as cotas necessárias ou entrar em contato com o AWS Support com antecedência.

Benefícios do estabelecimento desta prática recomendada: Saber as cotas de serviço, os limites de controle de utilização da API e as restrições de design permite que você os inclua no projeto, na implementação e na operação da carga de trabalho.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Analise as cotas de serviço da AWS na documentação publicada e no Service Quotas.
  - AWS Service Quotas (anteriormente chamado de limites)
- Examine o código da implantação para determinar todos os serviços necessários à sua workload.
- Use o AWS Config para encontrar todos os serviços da AWS usados na sua Contas da AWS.
  - Recursos da AWS Config compatíveis com o AWS, tipos e relacionamentos de recursos

- Você também pode usar o AWS CloudFormation para determinar os recursos da AWS usados. Examine os recursos que foram criados no AWS Management Console ou por meio do comando list-stack-resources da CLI. Você também pode ver no próprio modelo os recursos configurados para implantação.
  - Visualize dados e recursos da pilha do AWS CloudFormation no AWS Management Console
  - AWS CLI para o CloudFormation: list-stack-resources
- Determine as cotas de serviço aplicáveis. Use as informações acessíveis programaticamente por meio do Trusted Advisor e do Service Quotas.

Recursos

Documentos relacionados:

- AWS Marketplace: produtos CMDB que ajudam a acompanhar os limites
- AWS Service Quotas (anteriormente chamado de limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Limites de serviço)
- AWS Limit Monitor em AWS Answers
- Amazon EC2 Service Limits
- O que é o Service Quotas?

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP02 Gerenciar cotas de serviço de várias contas e regiões

Se você estiver usando várias Contas da AWS ou Regiões da AWS, solicite as cotas adequadas em todos os ambientes nos quais suas workloads de produção são executadas.

Cotas de serviço são rastreadas por conta. A menos que especificado de outra forma, cada cota é específica da Região da AWS. Além dos ambientes de produção, gerencie também as cotas em todos os ambientes aplicáveis que não são de produção, para que os testes e o desenvolvimento não sejam dificultados.

Antipadrões comuns:

- Permitir que a utilização de recursos em uma zona de isolamento cresça sem nenhum mecanismo para manter a capacidade das demais.
- Configurar manualmente todas as cotas nas zonas de isolamento independentemente.
- Não garantir que as implantações isoladas regionalmente sejam dimensionadas para acomodar o aumento do tráfego de outra região se uma implantação for perdida.

Benefícios do estabelecimento dessa prática recomendada: Ao assegurar o processamento da carga atual se uma zona de isolamento estiver indisponível, você ajuda a reduzir o número de erros ocorridos durante o failover, em vez de causar uma negação de serviço aos seus clientes.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Selecione as contas e as regiões relevantes conforme seus requisitos de serviço, de latência, regulatórios e de recuperação de desastres (DR).
- Identifique as cotas de serviço de todas as contas, regiões e zonas de disponibilidade relevantes. O escopo dos limites é definido para conta e região.
- O que é o Service Quotas?

Recursos

Documentos relacionados:

- AWS Marketplace: produtos CMDB que ajudam a acompanhar os limites
- AWS Service Quotas (anteriormente chamado de limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Limites de serviço)
- AWS Limit Monitor em AWS Answers
- Amazon EC2 Service Limits
- O que é o Service Quotas?

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP03 Acomodar 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.

Alguns exemplos incluem largura de banda de rede, tamanho da carga do AWS Lambda, taxa de intermitência de controle para o API Gateway e conexões simultâneas de usuários com um cluster do Amazon Redshift.

Antipadrões comuns:

- Realizar benchmarking por um período muito curto, utilizando o limite de intermitência, mas esperando que o serviço seja executado nessa capacidade por períodos prolongados.
- Escolher um design que usa um recurso de um serviço por usuário ou cliente, sem saber que há restrições que causarão falha nesse design à medida que você escala.

Benefícios do estabelecimento dessa prática recomendada: O acompanhamento de cotas fixas nos serviços da AWS e de restrições em outras partes da workload, como restrições de conectividade, de endereço IP e de serviços de terceiros, permite detectar quando você está propenso a uma determinada cota e resolvê-la antes que seja excedida.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Esteja ciente das cotas e restrições fixas de serviço e arquitete de forma correspondente.
  - AWS Service Quotas

Recursos

Documentos relacionados:

- AWS Marketplace: produtos CMDB que ajudam a acompanhar os limites
- AWS Service Quotas (anteriormente chamado de limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Limites de serviço)
- AWS Limit Monitor em AWS Answers
- Amazon EC2 Service Limits

- O que é o Service Quotas?

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP04 Monitorar e gerenciar cotas

Avalie seu uso potencial e aumente suas cotas adequadamente, permitindo o crescimento planejado do uso.

Para serviços compatíveis, você pode gerenciar suas cotas por meio da configuração de alarmes do CloudWatch para monitorar o uso e alertá-lo sobre as cotas prestes a serem atingidas. Esses alarmes podem ser acionados a partir de cotas de serviço ou do Service Quotas. Você também pode usar filtros de métrica no CloudWatch Logs para pesquisar e extrair padrões nos logs para determinar se o uso está se aproximando dos limites de cota.

Antipadrões comuns:

- Configurar alarmes para quando o Service Quotas estiver sendo atingido, mas não tiver um processo de resposta a um alerta.
- Configurar alarmes apenas para serviços compatíveis com o Service Quotas e não monitorar outros serviços.

Benefícios do estabelecimento dessa prática recomendada: O acompanhamento automático das cotas de serviço da AWS e o monitoramento do seu uso em relação a essas cotas permitirão que você veja quando estiver perto de atingir um limite. Você também pode usar esses dados de monitoramento para avaliar quando é possível reduzir cotas para economizar custos.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Monitore e gerencie cotas. Avalie seu uso potencial na AWS, aumente suas cotas de serviço regionais adequadamente e permita o crescimento planejado do uso.
  - Capture o consumo atual de recursos (por exemplo, buckets, instâncias). Use as operações de API de serviço, como a API DescribeInstances do Amazon EC2, para coletar o consumo atual de recursos.

- Capture suas cotas atuais. Use a documentação do AWS Service Quotas, AWS Trusted Advisor e da AWS.
  - Use o AWS Service Quotas é um serviço da AWS que ajuda você a gerenciar as cotas de mais de 100 serviços da AWS em um único local.
  - Use os limites de serviço do Trusted Advisor para determinar seus limites de serviço atuais.
  - Use as operações de API de serviço para determinar as cotas de serviço atuais, quando houver suporte.
  - Mantenha um registro dos aumentos de cota que foram solicitados e seus status. Após a aprovação de um aumento de cota, certifique-se de atualizar os registros para refletir a alteração.

Recursos

Documentos relacionados:

- AWS Marketplace: produtos CMDB que ajudam a acompanhar os limites
- AWS Service Quotas (anteriormente chamado de limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor para Limites de serviço
- AWS Limit Monitor em AWS Answers
- Amazon EC2 Service Limits
- O que é o Service Quotas?
- Monitoramento do Service Quotas com alarmes do Amazon CloudWatch

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP05 Automatizar o gerenciamento de cotas

Implemente ferramentas para alertar você quando os limites estiverem perto de serem atingidos. Ao usar as APIs do AWS Service Quotas, você pode automatizar as solicitações de aumento de cota.

Se você integrar o Configuration Management Database (CMDB) ou sistema de emissão de tíquetes com o Service Quotas, poderá automatizar o acompanhamento de solicitações de aumento de cota e

as cotas atuais. Além do AWS SDK, o Service Quotas oferece automação usando o AWS Command Line Interface (AWS CLI).

Antipadrões comuns:

- Acompanhar as cotas e o uso em planilhas.
- Executar relatórios sobre o uso diário, semanal ou mensal e comparar o uso com as cotas.

Benefícios do estabelecimento dessa prática recomendada: O acompanhamento automatizado das cotas de serviço da AWS e o monitoramento do seu uso em relação a essa cota permitem que você veja quando está perto de atingir um limite. Você pode configurar a automação para ajudá-lo a solicitar um aumento de cota quando necessário. Você pode considerar a redução de algumas cotas quando seu uso estiver na direção oposta para aproveitar os benefícios do risco reduzido (no caso de credenciais comprometidas) e da economia de custos.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Configure o monitoramento automatizado. Implemente ferramentas usando SDKs para alertar você quando os limites estiverem perto de serem atingidos.
  - Use o Service Quotas e aumente o serviço com uma solução automatizada de monitoramento de cotas, como o AWS Limit Monitor ou uma oferta do AWS Marketplace.
    - O que é o Service Quotas?
    - Monitoramento de cotas na AWS: solução da AWS
  - Configure respostas acionadas com base nos limites de cota por meio do Amazon SNS e das APIs do AWS Service Quotas.
  - Teste a automação.
    - Configure os limites.
    - Integre-se a eventos de alteração do AWS Config, de pipelines de implantação, do Amazon EventBridge ou de terceiros.
    - Defina limites baixos fictícios de cota para testar as respostas.
    - Configure gatilhos para executar a ação adequada mediante notificações e entre em contato com o AWS Support quando necessário.
    - Acione manualmente os eventos de alteração.
    - Execute um dia de jogo para testar o processo de alteração de aumento de cota.

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar no gerenciamento de configuração
- AWS Marketplace: produtos CMDB que ajudam a acompanhar os limites
- AWS Service Quotas (anteriormente chamado de limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Limites de serviço)
- Monitoramento de cotas na AWS: solução da AWS
- Amazon EC2 Service Limits
- O que é o Service Quotas?

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP06 Certificar-se de que 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 nas 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.

Antipadrões comuns:

- Configurar cotas de serviço com base nas necessidades atuais sem considerar os cenários de failover.

Benefícios do estabelecimento desta prática recomendada: Quando os eventos potencialmente afetam a disponibilidade, a nuvem permite que você implemente estratégias para atenuar ou recuperar estes eventos. Essas estratégias geralmente incluem a criação de recursos adicionais para substituir os que falharam. Sua estratégia de cota deve acomodar os recursos adicionais.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Certifique-se de que há uma lacuna suficiente entre a cota de serviço e o uso máximo para acomodar um failover.
  - Determine suas cotas de serviço, considerando os padrões de implantação, os requisitos de disponibilidade e o aumento do consumo.
  - Solicite aumentos de cota, se necessário. Planeje o tempo necessário para que as solicitações de aumento de cota sejam atendidas.
    - Determine os requisitos de confiabilidade (também conhecidos como número de noves).
    - Estabeleça os cenários de falha (por exemplo, perda de um componente, uma zona de disponibilidade ou uma região).
    - Estabeleça a metodologia de implantação (por exemplo, canário, azul/verde, vermelho/preto ou gradual).
    - Inclua um buffer adequado (por exemplo, 15%) do limite atual.
    - Planeje o aumento do consumo (por exemplo, monitore suas tendências de consumo).

Recursos

Documentos relacionados:

- AWS Marketplace: produtos CMDB que ajudam a acompanhar os limites
- AWS Service Quotas (anteriormente chamado de limites de serviço)
- Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Limites de serviço)
- Amazon EC2 Service Limits
- O que é o Service Quotas?

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas

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.

Práticas recomendadas

- REL02-BP01 Usar conectividade de rede altamente disponível nos endpoints públicos de workload
- REL02-BP02 Provisionar conectividade redundante entre as redes privadas na nuvem e nos ambientes on-premises
- REL02-BP03 Garantir contas de alocação de sub-rede IP para expansão e disponibilidade
- REL02-BP04 Preferir topologias hub-and-spoke em vez da malha muitos para muitos
- REL02-BP05 Aplicar intervalos de endereços IP privados não sobrepostos a todos os espaços de endereços privados onde estão conectados

REL02-BP01 Usar conectividade de rede altamente disponível nos endpoints públicos de workload

Esses endpoints e o roteamento para eles devem ser altamente disponíveis. Para que isso seja possível, use DNS altamente disponível, Redes de entrega de conteúdo (CDNs), API Gateway, balanceamento de carga ou proxies reversos.

O Amazon Route 53, a AWS o Global Accelerator, o Amazon CloudFront, o Amazon API Gateway, e o Elastic Load Balancing (ELB) fornecem endpoints públicos altamente disponíveis. Você também pode optar por avaliar os dispositivos de software do AWS Marketplace para o balanceamento de carga e o uso de proxy.

Os consumidores do serviço que sua carga de trabalho fornece, sejam eles usuários finais ou outros serviços, fazem solicitações nesses endpoints de serviço. Vários recursos da AWS estão disponíveis para permitir que você forneça endpoints altamente disponíveis.

O Elastic Load Balancing oferece balanceamento de carga entre zonas de disponibilidade, executa o roteamento da Camada 4 (TCP) ou da Camada 7 (http/https), integra-se ao AWS WAF e ao AWS Auto Scaling para ajudar a criar uma infraestrutura de autorreparação e absorver aumentos no tráfego com a liberação simultânea de recursos quando o tráfego diminuir.

O Amazon Route 53 é um serviço de Sistema de Nomes de Domínio (DNS) escalável e altamente disponível que conecta as solicitações de usuários à infraestrutura em execução na AWS, como instâncias do Amazon EC2, balanceadores de carga do Elastic Load Balancing ou buckets do Amazon S3. Além disso, também pode ser usado para direcionar os usuários para a infraestrutura fora da AWS.

O AWS Global Accelerator é um serviço de camada de rede que você pode usar para direcionar o tráfego para endpoints ideais pela rede global da AWS.

Ataques de negação de serviço distribuída (DDoS) arriscam interromper o tráfego legítimo e reduzir a disponibilidade para os seus usuários. O AWS Shield fornece proteção automática contra esses ataques, sem custo adicional para endpoints de serviços da AWS na sua workload. Expanda esses recursos com dispositivos virtuais de Parceiros do APN e o AWS Marketplace para atender às suas necessidades.

Antipadrões comuns:

- Usar endereços de Internet públicos em instâncias ou contêineres e gerenciar a conectividade com eles por meio de DNS.
- Usar endereços Internet Protocol em vez de nomes de domínio para localizar serviços.
- Fornecer conteúdo (páginas da web, ativos estáticos, arquivos de mídia) para uma grande área geográfica e não usar uma rede de entrega de conteúdo.

Benefícios do estabelecimento dessa prática recomendada: Com a implementação de serviços altamente disponíveis em sua carga de trabalho, você sabe que ela estará disponível aos seus usuários.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

Verifique se você tem conectividade altamente disponível para os usuários da workload. O Amazon Route 53, a AWS o Global Accelerator, o Amazon CloudFront, o Amazon API Gateway, e o Elastic Load Balancing (ELB) fornecem endpoints públicos altamente disponíveis. Você também pode optar por avaliar os dispositivos de software do AWS Marketplace para o balanceamento de carga e o uso de proxy.

- Verifique se você tem uma conexão altamente disponível para seus usuários.
- Verifique se você está usando um DNS altamente disponível para gerenciar os nomes de domínio dos endpoints da aplicação.
  - Se os usuários acessam seu aplicativo pela Internet, use as operações de API de serviço para confirmar o uso correto dos gateways da Internet. Confirme também se as entradas das tabelas de rotas para as sub-redes que hospedam os endpoints do seu aplicativo estão corretas.
    - DescribeInternetGateways

    - DescribeRouteTables
- Verifique se você está usando um proxy reverso ou um balanceador de carga altamente disponível na frente da aplicação.
    - Se os usuários acessam a aplicação por meio do ambiente on-premises, verifique se a conectividade entre a AWS e o ambiente é altamente disponível.
    - Use o Route 53 para gerenciar os nomes de domínio.
        - O que é o Amazon Route 53?
    - Use um provedor DNS de terceiros que atenda aos seus requisitos.
    - Use o Elastic Load Balancing.
        - O que é o Elastic Load Balancing?
    - Use um dispositivo do AWS Marketplace que atenda aos seus requisitos.

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar a planejar sua rede
- Recomendações de resiliência do AWS Direct Connect
- AWS Marketplace para infraestrutura de rede
- Whitepaper sobre as opções de conectividade do Amazon Virtual Private Cloud
- Multiple data center HA network connectivity
- Usar o Toolkit de resiliência do Direct Connect para começar
- VPC endpoints e serviços de VPC endpoint (AWS PrivateLink)
- O que é o AWS Global Accelerator?
- O que é o Amazon VPC?
- O que é um Transit Gateway?
- O que é o Amazon CloudFront?
- O que é o Amazon Route 53?
- O que é o Elastic Load Balancing?
- Trabalho com gateways Direct Connect

Vídeos relacionados:

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (NET406-R1)

REL02-BP02 Provisionar conectividade redundante entre as redes privadas na nuvem e nos ambientes on-premises

Use várias conexões do AWS Direct Connect ou túneis VPN entre as redes privadas implantadas separadamente. Use vários locais do Direct Connect 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.

O AWS Direct Connect é um serviço de nuvem que facilita a criação de uma conexão de rede dedicada entre seu ambiente on-premises e a AWS. Usando o Direct Connect Gateway, seu datacenter on-premises pode ser conectado a várias VPCs da AWS distribuídas em várias Regiões da AWS.

Essa redundância resolve possíveis falhas que afetam a resiliência da conectividade:

- Como você será resiliente a falhas em sua topologia?
- O que acontecerá se você configurar algo incorretamente e remover a conectividade?
- Você será capaz de lidar com um aumento inesperado no tráfego ou uso de seus serviços?
- Você conseguirá absorver uma tentativa de ataque de Negação de serviço distribuída (DDoS)?

Ao conectar sua VPC ao seu datacenter on-premises por meio de uma VPN, considere a resiliência e a largura de banda necessárias ao selecionar o fornecedor e o tamanho da instância em que precisa executar o dispositivo. Se você usar um dispositivo de VPN que não seja resiliente nesta implementação, precisará ter uma conexão redundante por meio de um segundo dispositivo. Para todos esses cenários, é preciso definir um tempo aceitável para recuperação e testar para garantir que você consiga cumprir esses requisitos.

Se você optar por conectar a VPC ao datacenter usando uma conexão Direct Connect e precisar que essa conexão seja altamente disponível, tenha conexões Direct Connect redundantes provenientes de cada datacenter. A conexão redundante deve usar uma segunda conexão Direct Connect de um local diferente do primeiro. Se você tiver vários datacenters, garanta que as conexões terminem em

diferentes locais. Use a ferramenta de recomendações do Toolkit de resiliência do Direct Connect para ajudar a configurar isso.

Se você escolher fazer failover para a VPN pela Internet usando a AWS VPN, saiba que ela é compatível com um throughput de até 1,25 Gbps por túnel VPN, mas não é compatível com Múltiplos caminhos de mesmo custo (ECMP) para tráfego de saída no caso de vários túneis da AWS Managed VPN terminarem no mesmo VGW. Não recomendamos que você use o AWS Managed VPN como backup para conexões Direct Connect, a menos que possa tolerar velocidades inferiores a 1 Gbps durante o failover.

Você também pode usar endpoints da VPC para conectar sua VPC a serviços compatíveis da AWS e do endpoint da VPC alimentado pelo AWS PrivateLink sem passar pela Internet pública. Os endpoints são dispositivos virtuais. Eles são componentes de VPC altamente disponíveis, redundantes e escalados horizontalmente. Eles permitem a comunicação entre instâncias em sua VPC e serviços sem impor riscos de disponibilidade ou restrições de largura de banda ao tráfego de rede.

Antipadrões comuns:

- Ter apenas um provedor de conectividade entre a rede local e a AWS.
- Consumir os recursos de conectividade da conexão do AWS Direct Connect, mas ter apenas uma conexão.
- Ter apenas um caminho para conectividade VPN.

Benefícios do estabelecimento dessa prática recomendada: Ao implementar conectividade redundante entre seu ambiente de nuvem e o ambiente corporativo ou on-premises, você pode garantir que os serviços dependentes entre os dois ambientes possam se comunicar de forma confiável.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Verifique se você tem conectividade altamente disponível entre a AWS e o ambiente on-premises. Use várias conexões do AWS Direct Connect ou túneis VPN entre as redes privadas implantadas separadamente. Use vários locais do Direct Connect 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.

- Verifique se você tem uma conexão redundante com seu ambiente on-premises. Você pode precisar de conexões redundantes para várias Regiões da AWS para atender às necessidades de disponibilidade.
  - Recomendações de resiliência do AWS Direct Connect
  - Uso de conexões Site-to-Site VPN redundantes para fornecer failover
    - Use as operações de API de serviço para identificar o uso correto dos circuitos do Direct Connect.
      - DescribeConnections
      - DescribeConnectionsOnInterconnect
      - DescribeDirectConnectGatewayAssociations
      - DescribeDirectConnectGatewayAttachments
      - DescribeDirectConnectGateways
      - DescribeHostedConnections
      - DescribeInterconnects
    - Se houver apenas uma conexão Direct Connect ou se você não tiver nenhuma, configure túneis VPN redundantes para seus gateways privados virtuais.
      - O que é a AWS Site-to-Site VPN?
- Capture a conectividade atual (por exemplo, Direct Connect, gateways privados virtuais, dispositivos do AWS Marketplace).
  - Use as operações de API de serviço para consultar a configuração das conexões Direct Connect.
    - DescribeConnections
    - DescribeConnectionsOnInterconnect
    - DescribeDirectConnectGatewayAssociations
    - DescribeDirectConnectGatewayAttachments
    - DescribeDirectConnectGateways
    - DescribeHostedConnections
    - DescribeInterconnects
  - Use as operações de API de serviço para coletar gateways privados virtuais onde as tabelas de rotas os usam.

    - DescribeVpnGateways
    - DescribeRouteTables
- Use as operações de API de serviço para coletar aplicações do AWS Marketplace onde as tabelas de rotas as utilizam.
    - DescribeRouteTables

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar a planejar sua rede
- Recomendações de resiliência do AWS Direct Connect
- AWS Marketplace para infraestrutura de rede
- Whitepaper sobre as opções de conectividade do Amazon Virtual Private Cloud
- Multiple data center HA network connectivity
- Uso de conexões Site-to-Site VPN redundantes para fornecer failover
- Usar o Toolkit de resiliência do Direct Connect para começar
- VPC endpoints e serviços de VPC endpoint (AWS PrivateLink)
- O que é o Amazon VPC?
- O que é um Transit Gateway?
- O que é a AWS Site-to-Site VPN?
- Trabalho com gateways Direct Connect

Vídeos relacionados:

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (NET406-R1)

REL02-BP03 Garantir contas de alocação de sub-rede IP para expansão e disponibilidade

Intervalos de endereços IP da Amazon VPC devem ser grandes o suficiente para acomodar os requisitos da workload, 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.

Ao planejar sua topologia de rede, a primeira etapa é definir o espaço do endereço IP em si. Intervalos de endereços IP privados (seguindo as diretrizes RFC 1918) devem ser alocados para cada VPC. Atenda aos seguintes requisitos como parte desse processo:

- Permitir espaço de endereço IP para mais de uma VPC por região.
- Dentro de uma VPC, deixe espaço para várias sub-redes que abrangem várias zonas de disponibilidade.
- Sempre deixe o espaço de bloco CIDR não utilizado em uma VPC para futura expansão.
- Verifique se há espaço de endereço IP para atender às necessidades de qualquer frota transitória de instâncias do EC2 que você use, como frotas spot para machine learning, clusters do Amazon EMR ou clusters do Amazon Redshift.
- Observe que os primeiros quatro endereços IP e o último endereço IP em cada bloco CIDR da sub-rede estão reservados e não estão disponíveis para seu uso.
- Você deve planejar implantar grandes blocos CIDR de VPC. Observe que o bloco CIDR inicial da VPC alocado para sua VPC não pode ser alterado ou excluído, mas você pode adicionar blocos CIDR não sobrepostos à VPC. Os CIDRs IPv4 da sub-rede não podem ser alterados, mas os CIDRs IPv6 podem. Lembre-se de que implantar a maior VPC possível (/16) resulta em mais de 65 mil endereços IP. Somente no espaço de endereço IP 10.x.x.x, você pode provisionar 255 dessas VPCs. Portanto, você deve errar por ser muito grande em vez de muito pequeno para facilitar o gerenciamento de suas VPCs.

Antipadrões comuns:

- Criar VPCs pequenas.
- Criar sub-redes pequenas e ter de adicionar sub-redes às configurações à medida que você cresce.
- Estimar incorretamente quantos endereços IP um Elastic Load Balancer pode usar.
- Implantar muitos load balancers de alto tráfego nas mesmas sub-redes.

Benefícios do estabelecimento dessa prática recomendada: Isso garante que você possa acomodar o crescimento das suas cargas de trabalho e continuar a fornecer disponibilidade à medida que elas se expandem.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Planeje sua rede para acomodar crescimento, conformidade regulamentar e integração com outras pessoas. O crescimento pode ser subestimado, a conformidade regulamentar pode mudar e as aquisições ou conexões de rede privada podem ser difíceis de implementar sem o planejamento adequado.
  - Selecione as Contas da AWS e regiões relevantes conforme seus requisitos de serviço, de latência, regulatórios e de recuperação de desastres (DR).
  - Identifique suas necessidades para implantações regionais de VPC.
  - Identifique o tamanho das VPCs.
    - Determine se você pretende implantar conectividade com várias VPCs.
      - O que é um Transit Gateway?
      - Conectividade com várias VPCs de região única
    - Determine se você precisa de rede segregada por requisitos regulamentares.
    - Faça VPCs o maior possível. O bloco CIDR inicial da VPC alocado para sua VPC não pode ser alterado ou excluído, mas você pode adicionar outros blocos CIDR não sobrepostos à VPC. No entanto, isso pode fragmentar seus intervalos de endereços.
    - Faça VPCs o maior possível. O bloco CIDR inicial da VPC alocado para sua VPC não pode ser alterado ou excluído, mas você pode adicionar outros blocos CIDR não sobrepostos à VPC. No entanto, isso pode fragmentar seus intervalos de endereços.

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar a planejar sua rede
- AWS Marketplace para infraestrutura de rede
- Whitepaper sobre as opções de conectividade do Amazon Virtual Private Cloud
- Multiple data center HA network connectivity
- Conectividade com várias VPCs de região única
- O que é o Amazon VPC?

Vídeos relacionados:

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (NET406-R1)

REL02-BP04 Preferir 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 on-premises) 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.

Se você tiver apenas duas redes desse tipo, basta conectá-las uma à outra, mas à medida que o número de redes cresce, a complexidade dessas conexões de malha torna-se insustentável. O AWS Transit Gateway oferece um modelo hub-and-spoke fácil de manter, permitindo o roteamento de tráfego em várias redes.

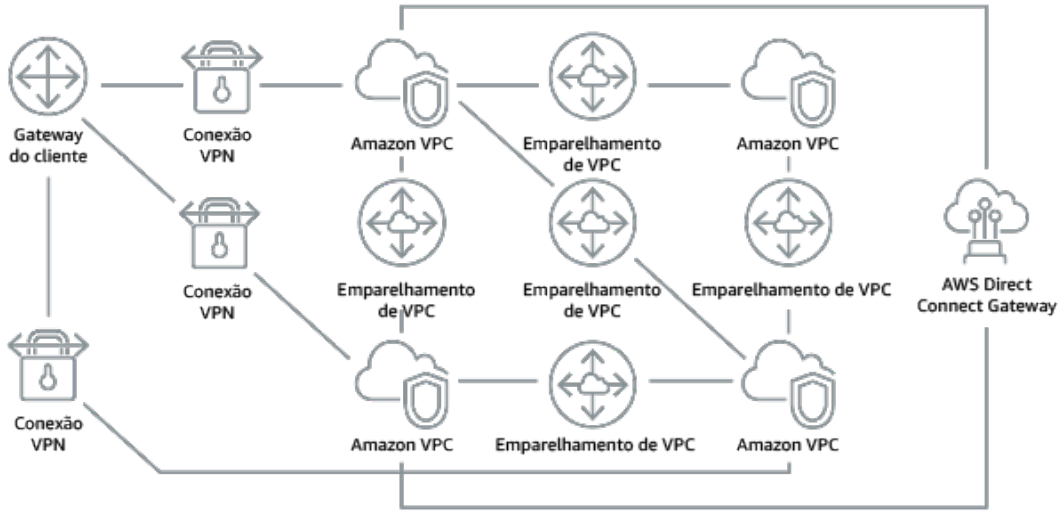


Figura 1: Sem o AWS Transit Gateway: você precisa emparelhar cada Amazon VPC com a outra e com cada localidade usando uma conexão VPN, que pode se tornar complexa à medida que ela escala.

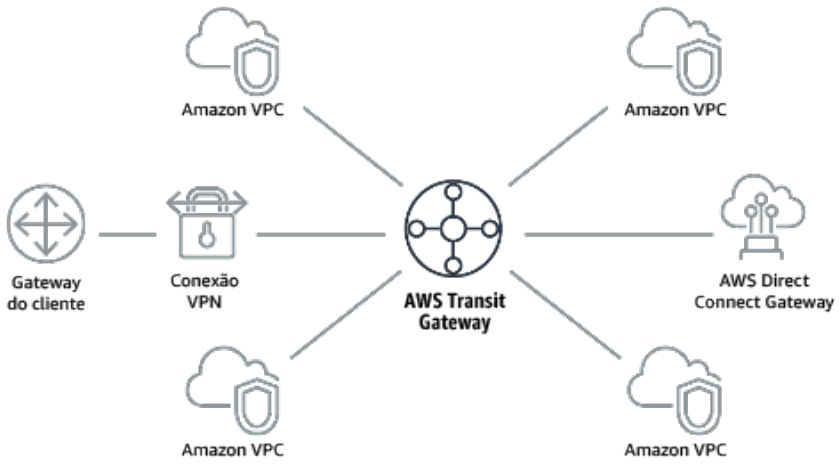


Figura 2: Com o AWS Transit Gateway: basta conectar cada Amazon VPC ou VPN ao AWS Transit Gateway e ele roteia o tráfego de e para cada VPC ou VPN.

Antipadrões comuns:

- Usar o emparelhamento de VPC para conectar mais de duas VPCs.
- Estabelecer várias sessões de BGP a cada VPC para fornecer conectividade que abrange as nuvens privadas virtuais (VPCs) distribuídas em diversas Regiões da AWS.

Benefícios do estabelecimento dessa prática recomendada: À medida que o número de redes cresce, a complexidade dessas conexões em malha torna-se insustentável. O AWS Transit Gateway oferece um modelo hub-and-spoke fácil de manter, que permite o roteamento do tráfego entre várias redes.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- 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 on-premises) 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.
  - Para apenas duas redes desse tipo, você pode simplesmente conectá-las uma à outra. No entanto, à medida que o número de redes cresce, a complexidade dessas conexões em malha

torna-se insustentável. O AWS Transit Gateway oferece um modelo hub-and-spoke fácil de manter, que permite o roteamento do tráfego entre várias redes.

- O que é um Transit Gateway?

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar a planejar sua rede
- AWS Marketplace para infraestrutura de rede
- Multiple data center HA network connectivity
- VPC endpoints e serviços de VPC endpoint (AWS PrivateLink)
- O que é o Amazon VPC?
- O que é um Transit Gateway?

Vídeos relacionados:

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (NET406-R1)

REL02-BP05 Aplicar intervalos de endereços IP privados não sobrepostos a todos os espaços de endereços privados onde 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.

Um sistema de gerenciamento de endereços IP (IPAM) pode ajudar com isso. Vários IPAMs estão disponíveis no AWS Marketplace.

Antipadrões comuns:

- Usar o mesmo intervalo de IPs na VPC que você tem no local ou na rede corporativa.
- Não acompanhar os intervalos IPs das VPCs usadas para implantar suas cargas de trabalho.

Benefícios do estabelecimento dessa prática recomendada: O planejamento ativo da rede garantirá que você não tenha várias ocorrências do mesmo endereço IP nas redes interconectadas. Isso evita que problemas de roteamento ocorram em partes da carga de trabalho que usam os diferentes aplicativos.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Monitore e gerencie seu uso do CIDR. Avalie seu uso potencial na AWS, adicione intervalos de CIDR às VPCs existentes e crie VPCs para permitir um crescimento planejado no uso.
  - Capture o consumo atual do CIDR (por exemplo, VPCs, sub-redes etc.)
    - Use as operações de API de serviço para coletar o consumo atual do CIDR.
  - Capture o seu uso atual de sub-rede.
    - Use as operações de API de serviço para coletar sub-redes por VPC em cada região.
      - DescribeSubnets
    - Registre o uso atual.
    - Determine se você criou algum intervalos de IP sobrepostos.
    - Calcule a capacidade não utilizada.
    - Identifique intervalos de IP sobrepostos. Você pode migrar para um novo intervalo de endereços ou usar os dispositivos de tradução de rede e porta (NAT) do AWS Marketplace se precisar conectar os intervalos sobrepostos.

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar a planejar sua rede
- AWS Marketplace para infraestrutura de rede
- Whitepaper sobre as opções de conectividade do Amazon Virtual Private Cloud
- Multiple data center HA network connectivity
- O que é o Amazon VPC?
- O que é o IPAM?

Vídeos relacionados:

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (NET406-R1)

Arquitetura da carga de trabalho

Perguntas

- REL 3 Como você projeta sua arquitetura de serviços de carga de trabalho?
- REL 4 Como você projeta interações em um sistema distribuído para evitar falhas?
- REL 5 Como você projeta interações em um sistema distribuído para mitigar ou resistir a falhas?

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.

Práticas recomendadas

- REL03-BP01 Escolher como segmentar a workload
- REL03-BP02 Criar serviços voltados a domínios e funcionalidades de negócios específicos
- REL03-BP03 Fornecer contratos de serviço por API

REL03-BP01 Escolher como segmentar a workload

A segmentação de workloads é importante ao determinar os requisitos de resiliência de sua aplicação. Uma arquitetura monolítica deve ser evitada sempre que possível. Em vez disso, considere cuidadosamente quais componentes da aplicação podem ser distribuídos em microsserviços. Dependendo dos requisitos de sua aplicação, isso pode acabar sendo uma combinação de uma arquitetura orientada a serviços (SOA) com microsserviços sempre que possível. Workloads com capacidade para serem do tipo sem estado têm maior chance de serem implantadas como microsserviços.

Resultado desejado: as workloads devem ser compatíveis, escaláveis e o mais vagamente agrupadas possível.

Ao tomar decisões sobre como segmentar uma workload, pondere os benefícios e as complexidades. O que é ideal para um novo produto a caminho do seu primeiro lançamento não se aplica a uma workload que foi criada para escalabilidade a partir das necessidades iniciais. Ao refatorar um monólito existente, você vai precisar considerar o quanto a aplicação vai oferecer um bom suporte a uma decomposição em direção à condição sem estado. A divisão dos serviços em pedaços menores permite que equipes pequenas e bem definidas os desenvolvam e gerenciem. No entanto, serviços menores podem introduzir complexidades que incluem maior latência potencial, depuração mais complexa e carga operacional aumentada.

Antipadrões comuns:

- O microsserviço Death Star é uma situação em que os componentes atômicos se tornam tão altamente interdependentes que a falha de um resulta em uma falha muito maior, o que torna os componentes tão rígidos e frágeis quanto um monólito.

Benefícios do estabelecimento desta prática:

- Mais segmentos específicos geram maior agilidade, flexibilidade organizacional e escalabilidade.
- Redução do impacto das interrupções do serviço.
- Os componentes da aplicação podem ter requisitos de disponibilidade diferentes, aos quais uma segmentação mais atômica pode oferecer suporte.
- Responsabilidades bem definidas para as equipes que oferecem suporte à workload.

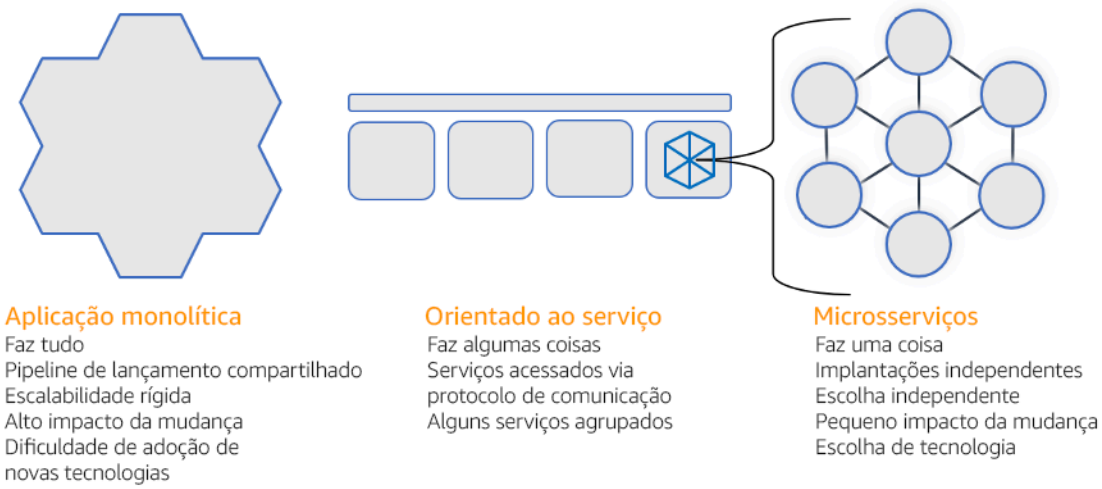
Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação de implementação

Escolha o tipo de arquitetura com base no modo como você segmentará a workload. Escolha uma SOA ou arquitetura de microsserviços (ou, em alguns casos, uma arquitetura monolítica). Mesmo que você opte por começar com uma arquitetura monolítica, você deve garantir que ela seja modular e tenha a capacidade de evoluir para SOA ou microsserviços à medida que o produto escala com a adoção do usuário. A SOA e os microsserviços oferecem, respectivamente, segmentação menor, que é preferida como uma arquitetura moderna escalável e confiável, mas há compensações a serem consideradas, especialmente ao implantar uma arquitetura de microsserviços.

Uma compensação primária é que você agora tem uma arquitetura de computação distribuída que pode tornar mais difícil alcançar requisitos de latência do usuário final, e há complexidade adicional na depuração e no rastreamento de interações com o usuário. Use o AWS X-Ray para

ajudar você a resolver esse problema. Outro efeito a ser considerado é o aumento da complexidade operacional à medida que você aumenta o número de aplicações que está gerenciando, o que requer a implantação de vários componentes de independência.



Arquiteturas monolítica, orientada a serviços e de microsserviços

Etapas da implementação

- Determine a arquitetura adequada para refatorar ou desenvolver sua aplicação. A SOA e os microsserviços oferecem respectivamente segmentação menor, que é preferida por ser uma arquitetura moderna escalável e confiável. A SOA pode ser o meio-termo ideal para alcançar uma segmentação menor e também evitar algumas das complexidades dos microsserviços. Para obter mais detalhes, consulte Compensações de microsserviços.
- Se sua carga de trabalho aceitá-la e sua organização puder sustentá-la, use uma arquitetura de microsserviços para obter a melhor agilidade e confiabilidade. Para obter mais detalhes, consulte Implementação de microsserviços na AWS.
- Considere seguir o padrão Strangler Fig para refatorar um monólito em componentes menores. Isso envolve a substituição gradual de componentes específicos da aplicação por novas aplicações e serviços. AWS Migration Hub Refactor Spaces atua como um ponto de partida para refatoração incremental. Para obter mais detalhes, consulte Migração simplificada de workloads on-premises herdadas usando um padrão strangler.
- A implementação de microsserviços pode exigir um mecanismo de descoberta de serviços para permitir que esses serviços distribuídos se comuniquem entre si. AWS App Mesh pode ser usado com arquiteturas orientadas por serviços para fornecer descoberta confiável e acesso a serviços. AWS Cloud Map também pode ser usado para descoberta dinâmica de serviços baseada em DNS.

- Se você estiver migrando de um monólito para SOA, Amazon MQ pode ajudar a eliminar a lacuna como um barramento de serviço ao reprojetar aplicações herdadas na nuvem.
- Para monólitos existentes com um único banco de dados compartilhado, escolha como reorganizar os dados em segmentos menores. Isso pode acontecer por unidade de negócios, padrão de acesso ou estrutura de dados. A esta altura no processo de refatoração, escolha se deseja prosseguir com um banco de dados relacional ou não relacional (NoSQL). Para obter mais detalhes, consulte De SQL para NoSQL.

Nível de esforço do plano de implementação: Alto

Recursos

Práticas recomendadas relacionadas:

- REL03-BP02 Criar serviços voltados a domínios e funcionalidades de negócios específicos

Documentos relacionados:

- Amazon API Gateway: configurar uma API REST usando o OpenAPI
- O que é arquitetura orientada a serviços?
- Contexto delimitado (um padrão central no design orientado por domínio)
- Implementação de microsserviços na AWS
- Compensações de microsserviços
- Microsserviços - uma definição desse novo termo de arquitetura
- Microsserviços na AWS
- O que é o AWS App Mesh?

Exemplos relacionados:

- Workshop de modernização iterativa de aplicações

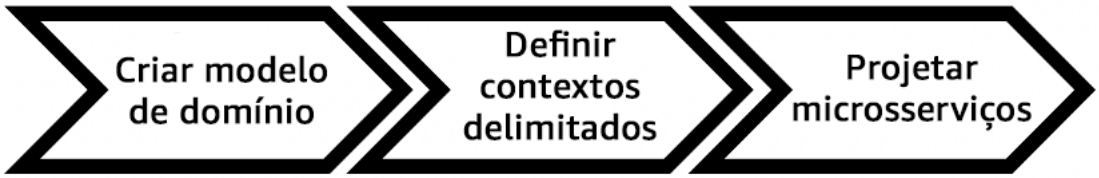
Vídeos relacionados:

- Delivering Excellence with Microservices on AWS (Entregando excelência com microsserviços na AWS)

REL03-BP02 Criar serviços voltados a domínios e funcionalidades de negócios específicos

A arquitetura orientada por serviços (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.

Ao projetar uma arquitetura de microsserviços, é útil usar o Design orientado por domínio (DDD) para modelar o problema de negócios usando entidades. Por exemplo, para o site Amazon.com, entidades podem incluir pacote, entrega, programação, preço, desconto e moeda. Em seguida, o modelo é dividido em modelos menores usando o Contexto delimitado, onde entidades que compartilham recursos e atributos semelhantes são agrupadas. Portanto, usar o pacote, a entrega e a programação de exemplo da Amazon.com seria parte do contexto de envio, enquanto preço, desconto e moeda fazem parte do contexto de definição de preço. Com o modelo dividido em contextos, surge um modelo de como delimitar microsserviços.



Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Projete a workload de acordo com os domínios de negócios e as respectivas funcionalidades. 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.
  - Execute a análise de domínio para mapear um Domain-Driven Design (DDD – Projeto orientado por domínio) para sua carga de trabalho. Em seguida, você pode escolher um tipo de arquitetura para atender às necessidades da sua workload.
    - Como dividir uma monolítica em microsserviços
    - Conceitos básicos do DDD quando cercado por sistemas herdados
    - Eric Evans “Design Orientado por Domínio: Lidando com a Complexidade no Coração do Software”

    - Implementação de microsserviços na AWS
- Decomponha os serviços nos menores componentes possíveis. Com a arquitetura de microsserviços, você pode separar sua carga de trabalho em componentes com a funcionalidade mínima para permitir escalabilidade e agilidade organizacionais.
  - Defina a API para a carga de trabalho e os respectivos objetivos, limites e outras considerações de uso do projeto.
    - Defina a API.
      - A definição da API deve permitir o crescimento e parâmetros adicionais.
    - Defina as disponibilidades projetadas.
      - Sua API pode ter vários objetivos de projeto para recursos diferentes.
    - Estabeleça limites
      - Use o teste para definir os limites de seus recursos de carga de trabalho.

Recursos

Documentos relacionados:

- Amazon API Gateway: configurar uma API REST usando o OpenAPI
- Contexto delimitado (um padrão central no design orientado por domínio)
- Eric Evans “Design Orientado por Domínio: Lidando com a Complexidade no Coração do Software”
- Conceitos básicos do DDD quando cercado por sistemas herdados
- Como dividir uma monolítica em microsserviços
- Implementação de microsserviços na AWS
- Compensações de microsserviços
- Microsserviços - uma definição desse novo termo de arquitetura
- Microsserviços na AWS

REL03-BP03 Fornecer 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 suas aplicações para a API mais recente quando estiverem prontas. 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.

Os microsserviços levam o conceito de arquitetura orientada a serviços (SOA) ao ponto de criar serviços com um conjunto mínimo de funcionalidades. Cada serviço publica uma API e projeta metas, limites e outras considerações para ele ser utilizado. Isso estabelece um contrato com chamadas a aplicações. Assim, três benefícios principais são alcançados:

- O serviço tem um problema de negócios conciso a ser resolvido e uma equipe pequena responsável por ele. Isso possibilita melhor escalabilidade organizacional.
- A equipe pode implantar a qualquer momento, desde que atenda aos requisitos de API e a outros requisitos do contrato.
- A equipe pode usar qualquer pilha de tecnologia desejada, desde que atenda os requisitos de API e outros requisitos do contrato.

O Amazon API Gateway é um serviço totalmente gerenciado que permite aos desenvolvedores criar, publicar, manter, monitorar e proteger APIs em qualquer escala com facilidade. Ele administra todas as tarefas envolvidas no recebimento e processamento de até centenas de milhares de chamadas de API simultâneas, inclusive gerenciamento de tráfego, controle de autorização e acesso, monitoramento, e gerenciamento de versões de API. Usando o OpenAPI Specification (OAS), anteriormente conhecido como Swagger Specification, você pode definir seu contrato de API e importá-lo para o API Gateway. Com o API Gateway, você pode controlar a versão e implantar as APIs.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Forneça contratos de serviço por API. Contratos de serviço são contratos documentados entre equipes na integração de serviços e incluem uma definição de API legível por máquina, limites de taxa e expectativas de performance.
  - Amazon API Gateway: configurar uma API REST usando o OpenAPI
    - 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.
    - O Amazon API Gateway é um serviço totalmente gerenciado que facilita a criação de APIs em qualquer escala para os desenvolvedores. Ao usar o OpenAPI Specification (OAS),

anteriormente conhecido como Swagger Specification, você pode definir seu contrato de API e importá-lo para o API Gateway. Com o API Gateway, você pode controlar a versão e implantar as APIs.

Recursos

Documentos relacionados:

- Amazon API Gateway: configurar uma API REST usando o OpenAPI
- Contexto delimitado (um padrão central no design orientado por domínio)
- Implementação de microsserviços na AWS
- Compensações de microsserviços
- Microsserviços - uma definição desse novo termo de arquitetura
- Microsserviços na AWS

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).

Práticas recomendadas

- REL04-BP01 Identificar qual tipo de sistema distribuído é necessário
- REL04-BP02 Implementar dependências com acoplamento fraco
- REL04-BP03 Fazer um trabalho constante
- REL04-BP04 Fazer com que todas as respostas sejam idempotentes

REL04-BP01 Identificar 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.

Os desafios mais difíceis com sistemas distribuídos são para sistemas complexos distribuídos em tempo real, também conhecidos como serviços de solicitação/resposta. O que as dificulta é que as solicitações chegam de forma imprevisível e as respostas devem ser fornecidas rapidamente (por exemplo, o cliente está aguardando ativamente a resposta). Os exemplos incluem servidores Web front-end, pipeline de pedidos, transações de cartão de crédito, todas as APIs da AWS e telefonia.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Identifique qual tipo de sistema distribuído é necessário. Os desafios dos sistemas distribuídos envolviam latência, escalabilidade, conhecimento das APIs de rede, marshalling e unmarshalling de dados e complexidade de algoritmos, como Paxos. À medida que os sistemas crescem e se tornam mais distribuídos, o que antes eram casos de borda hipotéticos se tornam ocorrências regulares.
  - A Amazon Builders' Library: desafios com sistemas distribuídos
    - Os sistemas distribuídos em tempo real rígidos exigem respostas síncronas e rápidas.
    - 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.

Recursos

Documentos relacionados:

- Amazon EC2: como garantir a idempotência
- A Amazon Builders' Library: desafios com sistemas distribuídos
- A Amazon Builders' Library: confiabilidade, trabalho constante e uma boa xícara de café
- O que é o Amazon EventBridge?
- O que é o Amazon Simple Queue Service?

Vídeos relacionados:

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Close Loops & Opening Minds: How to Take Control of Systems, Big & Small ARC337 (inclui acoplamento fraco, trabalho constante, estabilidade estática)
- AWS re:Invent 2019: Moving to event-driven architectures (SVS308)

REL04-BP02 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 de outros componentes que dependem dele, aumentando a resiliência e a agilidade.

Se as alterações em um componente forçarem outros componentes que dependem dele a serem também alterados, eles serão fortemente acoplados. O baixo acoplamento interrompe essa dependência para que os componentes dependentes só precisem saber a interface versionada e publicada. A implementação de um baixo acoplamento entre dependências isola uma falha em uma dependência para não afetar a outra.

O baixo acoplamento permite adicionar mais código ou recursos a um componente enquanto minimiza o risco para componentes que dependem dele. Além disso, a escalabilidade é melhorada pois você pode aumentar a escala verticalmente ou até mesmo alterar a implementação básica da dependência.

Para melhorar ainda mais a resiliência por meio do baixo acoplamento, torne as interações de componentes assíncronas sempre que possível. Esse modelo é adequado para qualquer interação que não precise de uma resposta imediata e em que uma confirmação de que uma solicitação foi registrada será suficiente. Envolve um componente que gera eventos e outro que os consome. Os dois componentes não se integram por meio de interação direta ponto a ponto, mas geralmente por meio de uma camada de armazenamento durável intermediária, como uma fila do SQS ou uma plataforma de dados de streaming, como o Amazon Kinesis ou o AWS Step Functions.

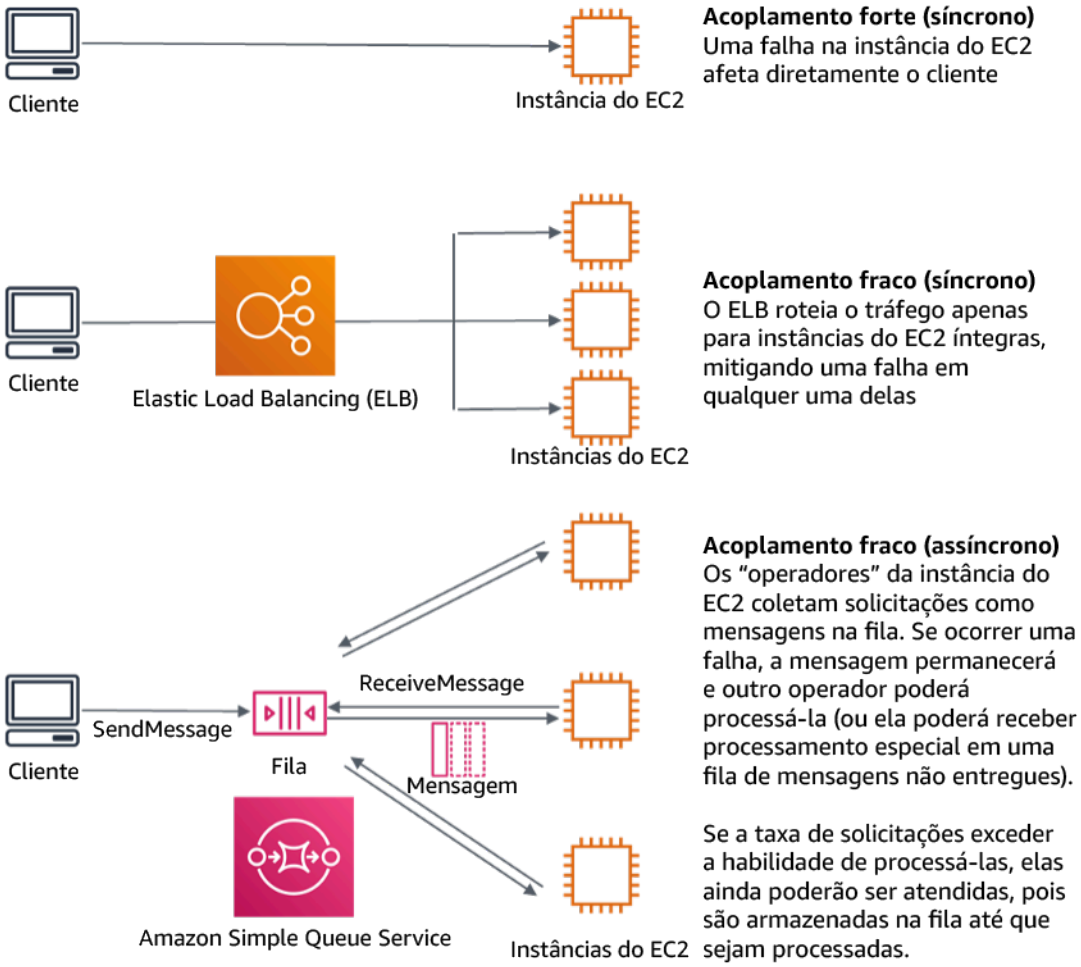


Figura 4: Dependências como sistemas de enfileiramento e load balancers têm baixo acoplamento

Filas do Amazon SQS e Elastic Load Balancers são apenas duas maneiras de adicionar uma camada intermediária para baixo acoplamento. Arquiteturas orientadas por eventos também podem ser criadas na Nuvem AWS usando o Amazon EventBridge, que pode abstrair clientes (produtores de eventos) dos serviços dos quais eles dependem (consumidores de eventos). O Amazon Simple Notification Service (Amazon SNS) é uma solução eficaz quando você precisa de mensagens de alto throughput, baseadas em push e de muitos para muitos. Usando tópicos do Amazon SNS, seus sistemas de editores podem enviar mensagens para um grande número de endpoints assinantes para processamento paralelo.

Embora as filas ofereçam várias vantagens, na maioria dos sistemas complexos em tempo real, as solicitações mais antigas do que um tempo limite (geralmente segundos) devem ser consideradas obsoletas (o cliente desistiu e não está mais esperando por uma resposta) e não processadas. Dessa forma, as solicitações mais recentes (e provavelmente ainda válidas) podem ser processadas.

Antipadrões comuns:

- Implantar um singleton como parte de uma carga de trabalho.
- Invocar diretamente as APIs entre níveis de carga de trabalho sem recurso de failover ou processamento assíncrono da solicitação.

Benefícios do estabelecimento desta prática recomendada: O baixo acoplamento ajuda a isolar o comportamento de um componente de outros componentes que dependem dele, aumentando a resiliência e a agilidade. A falha em um componente é isolada dos demais.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Implemente 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 de outros componentes que dependem dele, aumentando a resiliência e a agilidade.
  - AWS re:Invent 2019: Moving to event-driven architectures (SVS308)
  - O que é o Amazon EventBridge?
  - O que é o Amazon Simple Queue Service?
    - O Amazon EventBridge permite criar arquiteturas orientadas por eventos, que são acopladas de maneira fraca e distribuídas.
      - AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (MAD205)
    - Se as alterações em um componente forçarem outros componentes que dependem dele a serem também alterados, eles serão fortemente acoplados. O baixo acoplamento interrompe essa dependência para que os componentes dependentes precisem apenas reconhecer a interface versionada e publicada.
    - Sempre que possível, crie interações de componentes assíncronas. Esse modelo é adequado para qualquer interação que não precise de uma resposta imediata e quando uma confirmação de que uma solicitação foi registrada é suficiente.
      - AWS re:Invent 2019: Scalable serverless event-driven applications using Amazon SQS and Lambda (API304)

Recursos

Documentos relacionados:

- AWS re:Invent 2019: Moving to event-driven architectures (SVS308)
- Amazon EC2: como garantir a idempotência
- A Amazon Builders' Library: desafios com sistemas distribuídos
- A Amazon Builders' Library: confiabilidade, trabalho constante e uma boa xícara de café
- O que é o Amazon EventBridge?
- O que é o Amazon Simple Queue Service?

Vídeos relacionados:

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Close Loops & Opening Minds: How to Take Control of Systems, Big & Small ARC337 (inclui acoplamento fraco, trabalho constante, estabilidade estática)
- AWS re:Invent 2019: Moving to event-driven architectures (SVS308)
- AWS re:Invent 2019: Scalable serverless event-driven applications using Amazon SQS and Lambda (API304)

REL04-BP03 Fazer um trabalho constante

Os sistemas podem falhar quando há alterações grandes e rápidas na carga. Por exemplo, se a sua workload está realizando uma verificação de integridade que monitora a integridade de milhares de servidores, ela 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.

Por exemplo, se o sistema de verificação de integridade estiver monitorando 100.000 servidores, a carga nele será nominal a uma taxa de falha do servidor normalmente leve. No entanto, se um evento importante deixar metade desses servidores com problemas de integridade, o sistema de verificação de integridade ficará sobrecarregado tentando atualizar os sistemas de notificação e comunicar o estado com seus clientes. Portanto, em vez disso, o sistema de verificação de integridade deve enviar o snapshot completo do estado atual a cada vez. Os estados da integridade de 100.000 servidores, cada um representado por um bit, seriam apenas uma carga útil de 12,5 KB.

independentemente de nenhum servidor ou falhar, ou todos eles falharem, o sistema de verificação de integridade está realizando um trabalho constante, e alterações grandes e rápidas não são uma ameaça para a estabilidade do sistema. Na verdade, é assim que o Amazon Route 53 lida com as verificações de integridade de endpoints (como endereços IP) para determinar como os usuários finais são roteados para eles.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Faça um trabalho constante para que os sistemas não falhem quando houver mudanças rápidas e grandes na carga.
- Implemente 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 de outros componentes que dependem dele, aumentando a resiliência e a agilidade.
  - A Amazon Builders' Library: confiabilidade, trabalho constante e uma boa xícara de café
  - AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (inclui trabalho constante)
    - Para o exemplo de um sistema de verificação de integridade que monitora 100 mil servidores, crie as workloads de modo que os tamanhos da carga útil permaneçam constantes, seja qual for o número de êxitos ou falhas.

Recursos

Documentos relacionados:

- Amazon EC2: como garantir a idempotência
- A Amazon Builders' Library: desafios com sistemas distribuídos
- A Amazon Builders' Library: confiabilidade, trabalho constante e uma boa xícara de café

Vídeos relacionados:

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (inclui trabalho constante)

- AWS re:Invent 2018: Close Loops & Opening Minds: How to Take Control of Systems, Big & Small ARC337 (inclui acoplamento fraco, trabalho constante, estabilidade estática)
- AWS re:Invent 2019: Moving to event-driven architectures (SVS308)

REL04-BP04 Fazer 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.

Em um sistema distribuído, é fácil executar uma ação no máximo uma vez (o cliente faz apenas uma solicitação) ou pelo menos uma vez (continue solicitando até o cliente receber a confirmação do sucesso). Porém, é difícil garantir que uma ação seja idempotente, o que significa que ela é executada exatamente uma vez, de modo que fazer várias solicitações idênticas tenha o mesmo efeito de uma única solicitação. Usando tokens de idempotência em APIs, os serviços podem receber uma solicitação mutante uma vez ou mais sem a criação de registros duplicados nem efeitos colaterais.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- 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.
  - 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.
    - Amazon EC2: como garantir a idempotência

Recursos

Documentos relacionados:

- Amazon EC2: como garantir a idempotência
- A Amazon Builders' Library: desafios com sistemas distribuídos
- A Amazon Builders' Library: confiabilidade, trabalho constante e uma boa xícara de café

Vídeos relacionados:

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (MAD205)
- AWS re:Invent 2018: Close Loops & Opening Minds: How to Take Control of Systems, Big & Small ARC337 (inclui acoplamento fraco, trabalho constante, estabilidade estática)
- AWS re:Invent 2019: Moving to event-driven architectures (SVS308)

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.

Práticas recomendadas

- REL05-BP01 Implementar uma degradação simples para transformar dependências rígidas aplicáveis em dependências flexíveis
- REL05-BP02 Controlar o fluxo de solicitações
- REL05-BP03 Controlar e limitar as chamadas de repetição
- REL05-BP04 Antecipar-se à falha e filas limitadas
- REL05-BP05 Definir tempos limite do cliente
- REL05-BP06 Criar serviços sem estado sempre que possível
- REL05-BP07 Implementar medidas emergenciais

REL05-BP01 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.

Considere um serviço B que é chamado pelo serviço A e, por sua vez, chama o serviço C.

Figura 5: O serviço C falha quando chamado do serviço B. O serviço B retorna uma resposta degradada ao serviço A.

Quando o serviço B chama o serviço C, ele recebeu um erro ou tempo limite dele. O serviço B, sem uma resposta do serviço C (e os dados que ele contém), retorna o que pode. Esse pode ser o último bom valor armazenado em cache, ou o serviço B pode substituir uma resposta estática pré-determinada pelo que receberia do serviço C. Em seguida, ele pode retornar uma resposta degradada ao chamador, o serviço A. Sem essa resposta estática, a falha no serviço C seria feita em cascata por meio do serviço B para o serviço A, resultando em uma perda de disponibilidade.

De acordo com o fator multiplicativo na equação de disponibilidade para dependências rígidas (consulte Cálculo de disponibilidade com dependências rígidas), qualquer queda na disponibilidade do C afeta gravemente a disponibilidade efetiva do B. Ao retornar a resposta estática, o serviço B atenua a falha em C e, embora degradada, faz com que a disponibilidade do serviço C pareça 100% (supondo que ela retorne de forma confiável a resposta estática sob condições de erro). Observe que a resposta estática é uma alternativa simples para retornar um erro e não é uma tentativa de recalcular a resposta usando meios diferentes. Essas tentativas em um mecanismo completamente diferente para tentar alcançar o mesmo resultado são chamadas de comportamento de fallback e são um antipadrão a ser evitado.

Outro exemplo de degradação tranquila é o padrão de disjuntor. Estratégias de repetição devem ser usadas quando a falha é transitória. Quando esse não for o caso, e a operação provavelmente falhar, o padrão do disjuntor impedirá que o cliente execute uma solicitação que provavelmente falhará. Quando as solicitações estão sendo processadas normalmente, o disjuntor está fechado e as solicitações passam. Quando o sistema remoto começa a retornar erros ou exibe alta latência,

o disjuntor abre e a dependência é ignorada ou os resultados são substituídos por respostas mais simples, mas menos abrangentes (que podem ser simplesmente um cache de resposta). O sistema periodicamente tenta chamar a dependência para determinar se ela se recuperou. Quando isso acontece, o disjuntor é fechado.

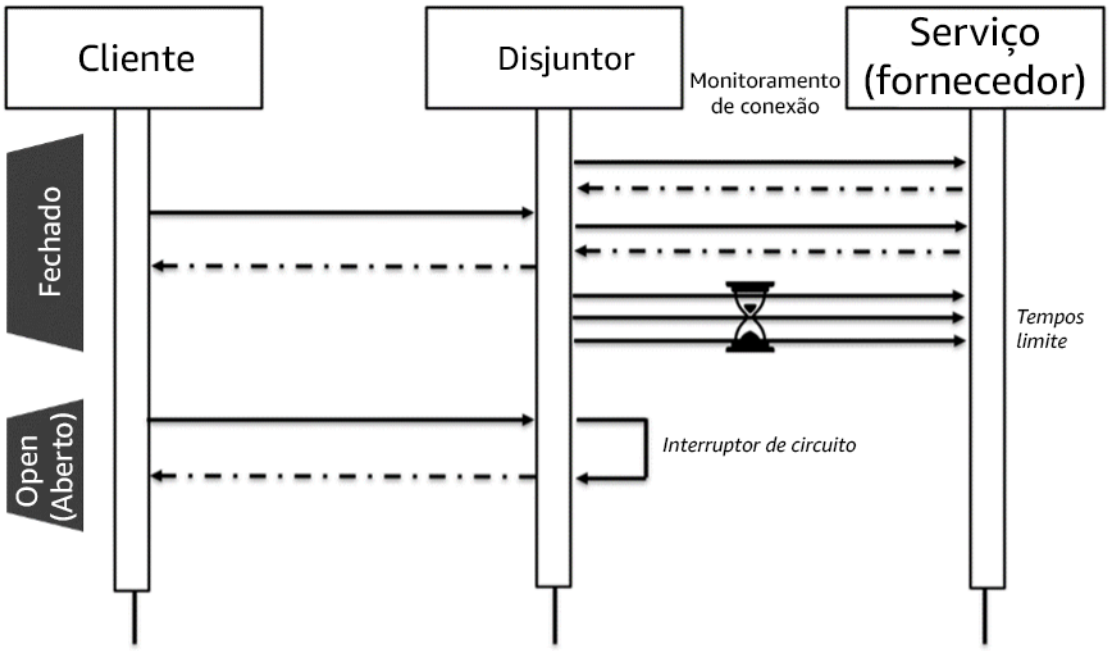


Figura 6: Disjuntor mostrando estados fechados e abertos.

Além dos estados fechado e aberto mostrados no diagrama, após um período configurável no estado aberto, o disjuntor pode fazer a transição para meio aberto. Nesse estado, ele tenta chamar o serviço periodicamente a uma taxa muito menor do que o normal. Esse teste é usado para verificar a integridade do serviço. Depois de vários êxitos no estado meio aberto, o disjuntor muda para fechado, e as solicitações normais são retomadas.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Implemente 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.
  - Ao retornar uma resposta estática, a workload atenua as falhas que ocorrem nas dependências dela.

    - Laboratório do Well-Architected: nível 300: implementação de verificações de integridade e do gerenciamento de dependências para melhorar a confiabilidade
- Detecte quando há probabilidade de falha na operação de repetição e impeça o cliente de fazer chamadas com falha com o padrão de disjuntor.
    - CircuitBreaker

Recursos

Documentos relacionados:

- Amazon API Gateway: controlar o fluxo de solicitações de API para uma melhor produtividade
- CircuitBreaker (resume “Circuit Breaker” do livro “Release It!”)
- Repetições de erros e recuo exponencial na AWS
- Michael Nygard “Release It! Design and Deploy Production-Ready Software”
- A Amazon Builders’ Library: evitar fallback em sistemas distribuídos
- A Amazon Builders’ Library: evitar backlogs de fila insuperáveis
- A Amazon Builders’ Library:desafios e estratégias de armazenamento em cache
- A Amazon Builders’ Library: tempos limite, novas tentativas e recuo com tremulação

Vídeos relacionados:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders’ Library (DOP328)

Exemplos relacionados:

- Laboratório do Well-Architected: nível 300: implementação de verificações de integridade e do gerenciamento de dependências para melhorar a confiabilidade

REL05-BP02 Controlar o fluxo de solicitações

O controle de utilização de solicitações é um padrão de atenuação 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.

Seus serviços devem ser projetados para processar uma capacidade conhecida de solicitações que cada nó ou célula pode processar. Esta capacidade pode ser estabelecida por meio de teste de carga. É preciso acompanhar a taxa de chegada das solicitações e, se ela ultrapassar esse limite, a resposta adequada será indicar que a solicitação foi limitada. Isso permite que o usuário tente outra vez, possivelmente para um nó ou célula diferente que talvez tenha capacidade disponível. O Amazon API Gateway fornece métodos para controle de solicitações. O Amazon SQS e o Amazon Kinesis podem armazenar solicitações em buffer, suavizar a taxa de solicitações e aliviar a necessidade de controle de utilização para solicitações que podem ser abordadas de forma assíncrona.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Controle o fluxo de solicitações. 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.
  - Use o Amazon API Gateway
    - Controlar o fluxo de solicitações de API para uma melhor produtividade

Recursos

Documentos relacionados:

- Amazon API Gateway: controlar o fluxo de solicitações de API para uma melhor produtividade
- Repetições de erros e recuo exponencial na AWS
- A Amazon Builders' Library: evitar fallback em sistemas distribuídos
- A Amazon Builders' Library: evitar backlogs de fila insuperáveis
- A Amazon Builders' Library: tempos limite, novas tentativas e recuo com tremulação
- Controlar o fluxo de solicitações de API para uma melhor produtividade

Vídeos relacionados:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)

REL05-BP03 Controlar e limitar 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.

Os componentes típicos em um sistema de software distribuído incluem servidores, load balancers, bancos de dados e servidores DNS. Em operação, e sujeito a falhas, qualquer um deles pode começar a gerar erros. A técnica padrão para lidar com erros é implementar novas tentativas no lado do cliente. Essa técnica aumenta a confiabilidade e a disponibilidade do aplicativo. No entanto, em grande escala (e se os clientes tentarem repetir a operação com falha assim que ocorrer um erro) a rede poderá ficar rapidamente saturada com solicitações novas e repetidas, cada uma competindo pela largura de banda da rede. Isso pode resultar em uma tempestade de repetições, o que reduzirá a disponibilidade do serviço. Esse padrão pode continuar até que ocorra uma falha completa do sistema.

Para evitar tais cenários, algoritmos de recuo, como o recuo exponencial comum, devem ser usados. Os algoritmos de recuo exponencial diminuem gradualmente a taxa na qual novas tentativas são realizadas, evitando assim congestionamentos de rede.

Muitos SDKs e bibliotecas de software, incluindo os da AWS, implementam uma versão desses algoritmos. No entanto, nunca presuma que exista um algoritmo de recuo, sempre teste e verifique se esse é o caso.

O recuo simples não é suficiente porque, em sistemas distribuídos, todos os clientes podem recuar simultaneamente, criando clusters de chamadas de repetição. Marc Brooker, em sua publicação de blog , Recuo exponencial e jitter, explica como modificar a função wait() no recuo exponencial para impedir clusters de chamadas de repetição. A solução é adicionar jitter na função wait(). Para evitar tentar novamente por muito tempo, as implementações devem limitar o recuo a um valor máximo.

Por fim, é importante configurar um número máximo de repetições ou tempo decorrido, após o qual uma repetição simplesmente falhará. Os AWS SDKs implementam isso por padrão, o que pode ser configurado. Para serviços mais baixos na pilha, um limite máximo de repetição zero ou um pode limitar o risco, mas ainda ser eficaz à medida que novas tentativas forem delegadas a serviços mais altos na pilha.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- 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.
  - Repetições de erros e recuo exponencial na AWS
    - Os SDKs da Amazon implementam repetições e recuo exponencial por padrão. Implemente uma lógica semelhante em sua camada de dependência ao chamar seus próprios serviços dependentes. Decida quais são os tempos limite e quando parar de tentar novamente com base no seu caso de uso.

Recursos

Documentos relacionados:

- Amazon API Gateway: controlar o fluxo de solicitações de API para uma melhor produtividade
- Repetições de erros e recuo exponencial na AWS
- A Amazon Builders' Library: evitar fallback em sistemas distribuídos
- A Amazon Builders' Library: evitar backlogs de fila insuperáveis
- A Amazon Builders' Library:desafios e estratégias de armazenamento em cache
- A Amazon Builders' Library: tempos limite, novas tentativas e recuo com tremulação

Vídeos relacionados:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)

REL05-BP04 Antecipar-se à falha 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.

Essa melhor prática se aplica ao lado do servidor, ou receptor, da solicitação.

Esteja ciente de que as filas podem ser criadas em vários níveis de um sistema e podem impedir seriamente a capacidade de recuperação rápida à medida que solicitações antigas obsoletas (que não precisam mais de uma resposta) são processadas antes de solicitações mais recentes. Esteja ciente dos locais onde as filas existem. Elas geralmente se ocultam em fluxos de trabalho ou em trabalhos registrados em um banco de dados.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Antecipe-se à falha e limite filas. 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.
  - Implemente antecipação à falha quando o serviço estiver sob pressão.
    - Falha rápida
  - Filas limitadas. Em um sistema baseado em fila, quando o processamento é interrompido, mas as mensagens continuam chegando, o débito de mensagens pode se acumular em uma lista grande de pendências, aumentando o tempo de processamento. Os resultados podem deixar de ser úteis por conta da demora na conclusão do trabalho, o que afeta principalmente a disponibilidade que o enfileiramento tinha que proteger.
    - A Amazon Builders' Library: evitar backlogs de fila insuperáveis

Recursos

Documentos relacionados:

- Repetições de erros e recuo exponencial na AWS
- Falha rápida
- A Amazon Builders' Library: evitar fallback em sistemas distribuídos
- A Amazon Builders' Library: evitar backlogs de fila insuperáveis
- A Amazon Builders' Library:desafios e estratégias de armazenamento em cache
- A Amazon Builders' Library: tempos limite, novas tentativas e recuo com tremulação

Vídeos relacionados:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)

REL05-BP05 Definir tempos limite do cliente

Defina tempos limite adequados, verifique-os sistematicamente e não dependa de valores padrão, já que eles costumam ser muito altos.

Essa melhor prática se aplica ao lado do cliente, ou remetente, da solicitação.

Defina um tempo limite de conexão e um tempo limite de solicitação em qualquer chamada remota e, normalmente, em qualquer chamada entre processos. Muitas estruturas de trabalho oferecem recursos de tempo limite integrados, mas tenha cuidado, porque muitos deles têm valores padrão infinitos ou muito altos. Um valor muito alto reduz a utilidade do tempo limite porque os recursos continuam a ser consumidos enquanto o cliente aguarda o decorrer do tempo limite. Um valor muito baixo pode gerar maior tráfego no back-end e maior latência, porque muitas solicitações são repetidas. Em alguns casos, isso pode levar a interrupções completas porque todas as solicitações estão sendo repetidas.

Para saber mais sobre como a Amazon usa tempos limite, repetições e recuo com tremulação, consulte a Builder's Library: tempos limite, repetições e recuo com tremulação.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Defina um tempo limite de conexão e um tempo limite de solicitação em qualquer chamada remota e, normalmente, em qualquer chamada entre processos. Muitas estruturas de trabalho oferecem recursos de tempo limite integrados, mas tenha cuidado, porque muitos deles têm valores padrão infinitos ou muito altos. Um valor muito alto reduz a utilidade do tempo limite porque os recursos continuam a ser consumidos enquanto o cliente aguarda o decorrer do tempo limite. Um valor muito baixo pode gerar maior tráfego no back-end e maior latência, porque muitas solicitações são repetidas. Em alguns casos, isso pode levar a interrupções completas porque todas as solicitações estão sendo repetidas.
  - AWS SDK: repetições e tempos limite

Recursos

Documentos relacionados:

- AWS SDK: repetições e tempos limite
- Amazon API Gateway: controlar o fluxo de solicitações de API para uma melhor produtividade
- Repetições de erros e recuo exponencial na AWS
- A Amazon Builders' Library: tempos limite, novas tentativas e recuo com tremulação

Vídeos relacionados:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)

REL05-BP06 Criar 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 causar impacto na disponibilidade. O Amazon ElastiCache ou o Amazon DynamoDB são bons destinos para o estado descarregado.

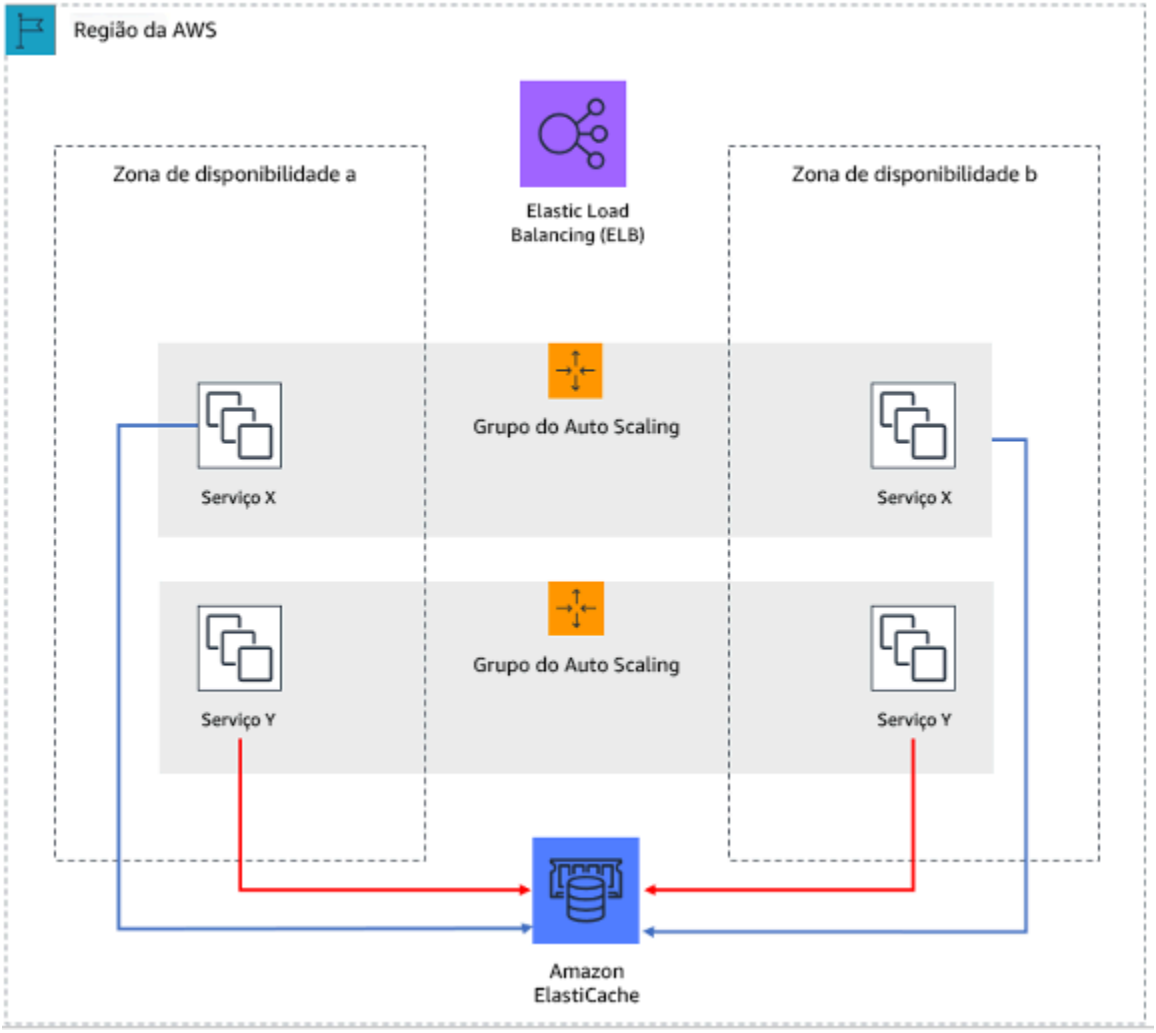


Figura 7: Nesta aplicação Web sem estado, o estado da sessão é descarregado para o Amazon ElastiCache.

Quando os usuários ou serviços interagem com um aplicativo, eles geralmente executam uma série de interações que formam uma sessão. Uma sessão são dados exclusivos para usuários que persistem entre solicitações enquanto usam o aplicativo. Um aplicativo sem estado é um aplicativo que não precisa de conhecimento de interações anteriores e não armazena informações da sessão.

Depois de projetados para serem sem estado, você pode usar serviços de computação com tecnologia sem servidor, como o AWS Lambda ou o AWS Fargate.

Além da substituição do servidor, outro benefício dos aplicativos sem estado é que eles podem escalar horizontalmente, pois qualquer um dos recursos de computação disponíveis (como instâncias do EC2 e funções do AWS Lambda) pode atender a qualquer solicitação.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Crie aplicações sem estado. Os aplicativos sem estado permitem a escalabilidade horizontal e são tolerantes a falhas de um nó individual.
  - Remova o estado que realmente pode ser armazenado nos parâmetros de solicitação.
  - Depois de examinar se o estado é necessário, mova qualquer rastreamento de estado para um armazenamento em cache resiliente multizona ou armazenamento de dados, como o Amazon ElastiCache, o Amazon RDS, Amazon DynamoDB ou uma solução de dados distribuídos de terceiros. Armazene os estados que não puderam ser movidos para armazenamentos de dados resilientes.
    - Alguns dados (como cookies) podem ser inseridos em cabeçalhos ou parâmetros de consulta.
    - Faça a refatoração para remover o estado que pode ser inserido rapidamente nas solicitações.
    - Alguns dados talvez não sejam realmente necessários por solicitação e podem ser recuperados sob demanda.
    - Remova os dados que podem ser recuperados de forma assíncrona.
    - Escolha um armazenamento de dados que atenda aos requisitos de um estado necessário.
    - Considere um banco de dados NoSQL para dados não relacionais.

Recursos

Documentos relacionados:

- A Amazon Builders' Library: evitar fallback em sistemas distribuídos
- A Amazon Builders' Library: evitar backlogs de fila insuperáveis
- A Amazon Builders' Library:desafios e estratégias de armazenamento em cache

REL05-BP07 Implementar medidas emergenciais

Medidas emergenciais são processos rápidos que podem atenuar o impacto da disponibilidade na workload.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Implemente 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. Geralmente, as medidas são manuais, mas também podem ser automatizadas
  - Exemplos de medidas incluem
    - Bloquear todo tráfego de robô
    - Servir páginas estáticas em vez de dinâmicas
    - Reduzir a frequência de chamadas a uma dependência
    - Limitar as chamadas de dependências
  - Dicas para implementar e usar medidas emergenciais
    - Quando as medidas forem ativadas, faça MENOS, e não mais
    - Simplifique, evite comportamento bimodal
    - Teste suas medidas periodicamente
  - Veja a seguir exemplos de ações que NÃO são medidas emergenciais
    - Adicionar capacidade
    - Chamar proprietários de serviços de clientes que dependem do seu serviço e solicitar que eles reduzam as chamadas
    - Fazer uma alteração no código e lançá-lo

Gerenciamento de alterações

Perguntas

- REL 6 Como você monitora recursos de carga de trabalho?
- REL 7 Como você projeta sua carga de trabalho para se adaptar às mudanças na demanda?
- REL 8 Como você implementa uma alteração?

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.

Práticas recomendadas

- REL06-BP01 Monitorar todos os componentes da workload (geração)
- REL06-BP02 Definir e calcular as métricas (agregação)
- REL06-BP03 Enviar notificações (processamento e emissão de alarmes em tempo real)
- REL06-BP04 Automatizar respostas (processamento e emissão de alarmes em tempo real)
- REL06-BP05 Análises
- REL06-BP06 Realizar revisões regularmente
- REL06-BP07 Monitorar o rastreamento completo das solicitações por meio do seu sistema

REL06-BP01 Monitorar todos os componentes da workload (geração)

monitore os componentes da carga de trabalho com o Amazon CloudWatch ou ferramentas de terceiros. Monitore os serviços da AWS com o painel do AWS Health.

Todos os componentes da carga de trabalho devem ser monitorados, incluindo front-end, lógica de negócios e níveis de armazenamento. Defina as principais métricas, descreva como extraí-las dos logs (se necessário) e defina limites de ativação para eventos de alarme correspondentes. Certifique-se de que as métricas sejam relevantes para os indicadores-chave de performance (KPIs) da workload e use métricas e logs para identificar os primeiros sinais de alerta de degradação do serviço. Por exemplo, uma métrica relacionada a resultados de negócios, como o número de pedidos processados com êxito por minuto, pode indicar problemas de workload mais rapidamente do que uma métrica técnica, como a utilização da CPU. Use o painel do AWS Health para uma visualização personalizada da performance e da disponibilidade dos serviços da AWS subjacentes aos recursos da AWS.

O monitoramento na nuvem oferece novas oportunidades. A maioria dos provedores de nuvem desenvolveu ganchos personalizáveis e pode entregar insights para ajudar você a monitorar várias camadas da workload. Serviços da AWS, como o Amazon CloudWatch, aplicam algoritmos estatísticos e de machine learning para analisar continuamente métricas de sistemas e de aplicações, determinam linhas de base normais e detectam anomalias com intervenção mínima do

usuário. Os algoritmos de detecção de anomalias consideram a sazonalidade e as mudanças de tendência das métricas.

A AWS disponibiliza uma abundância de informações de monitoramento e de log para consumo, que podem ser usadas para definir métricas específicas de workload, processos de alteração sob demanda e adotar técnicas de machine learning, independentemente da experiência em ML.

Além disso, monitore todos os seus endpoints externos para garantir que eles sejam independentes de sua implementação de base. Este monitoramento ativo pode ser feito com transações sintéticas (às vezes chamadas de canários de usuário, mas que não devem ser confundido com implantações canário) que executam periodicamente um número de tarefas comuns que correspondem às ações realizadas pelos clientes da workload. Mantenha estas tarefas de curta duração e certifique-se de não sobrecarregar a workload durante o teste. O Amazon CloudWatch Synthetics permite criar canários sintéticos para monitorar seus endpoints e APIs. Você também pode combinar os nós sintéticos do cliente canário com o console do AWS X-Ray para identificar quais canários sintéticos estão enfrentando problemas com erros, falhas ou taxas de controle de utilização para o período selecionado.

Resultado desejado:

Coletar e usar métricas críticas de todos os componentes da workload para garantir sua confiabilidade e a experiência ideal do usuário. Detectar que uma workload não está alcançando resultados de negócios permite que você declare rapidamente um desastre e se recupere de um incidente.

Antipadrões comuns:

- Monitorar apenas as interfaces externas com sua carga de trabalho.
- Não gerar métricas específicas de workload e confiar apenas nas métricas fornecidas pelos serviços da AWS usados pela sua workload.
- Usar apenas métricas técnicas na workload e não monitorar nenhuma métrica relacionada a KPIs não técnicos para os quais a workload contribui.
- Depender do tráfego de produção e de verificações de integridade simples para monitorar e avaliar o estado da workload.

Benefícios do estabelecimento dessa prática recomendada: O monitoramento em todos os níveis da workload permite prever e resolver problemas mais rapidamente nos componentes que a compõem.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

1. Habilite o registro em log quando disponível. Os dados de monitoramento devem ser obtidos de todos os componentes das workloads. Ative o registro em log adicional, como os logs de acesso do S3, e habilite sua workload para registrar dados específicos da workload. Colete métricas para médias de CPU, E/S de rede e E/S de disco de serviços como o Amazon ECS, o Amazon EKS, o Amazon EC2, o Elastic Load Balancing, o AWS Auto Scaling e o Amazon EMR. Perceber Serviços da AWS que publicam métricas do CloudWatch para uma lista dos serviços da AWS que publicam métricas do CloudWatch.
2. Revise todas as métricas padrão e explore quaisquer lacunas na coleta de dados. Cada serviço gera métricas padrão. A coleta de métricas padrão permite que você entenda melhor as dependências entre os componentes da workload e como a confiabilidade e a performance destes componentes a afetam. Você também pode criar e publicar suas próprias métricas para CloudWatch usando o AWS CLI ou uma API. Isso
3. Avalie todas as métricas para decidir quais alertar para cada serviço da AWS na sua workload. Você pode escolher selecionar um subconjunto de métricas que tenha um grande impacto na confiabilidade da workload. Focar em métricas e limites críticos permite refinar o número de alertas de emergência e pode ajudar a minimizar falso-positivos.
4. Defina alertas e o processo de recuperação para a workload depois que o alerta for acionado. A definição de alertas permite que você notifique, escalone e siga rapidamente as etapas necessárias para se recuperar de um incidente e atender ao seu objetivo de tempo de recuperação (RTO) prescrito. Você pode usar o alarmes do Amazon CloudWatch para invocar fluxos de trabalho automatizados e iniciar procedimentos de recuperação com base em limites definidos.
5. Explore o uso de transações sintéticas para coletar dados relevantes sobre o estado das workloads. O monitoramento sintético segue as mesmas rotas e realiza as mesmas ações que um cliente, possibilitado que você verifique continuamente a experiência do cliente, mesmo quando não há tráfego de clientes nas workloads. Ao usar transações sintéticas, você pode descobrir problemas antes que seus clientes o façam.

Recursos

Práticas recomendadas relacionadas:

- REL11-BP03 Automatizar a reparação em todas as camadas

Documentos relacionados:

- Conceitos básicos do painel do AWS Health: integridade da sua conta
- Serviços da AWS que publicam métricas do CloudWatch
- Logs de acesso para o Network Load Balancer
- Logs de acesso para seu application load balancer
- Acessar o Amazon CloudWatch Logs para o AWS Lambda
- Registro em log de acesso ao servidor do Amazon S3
- Habilite logs de acesso para o Classic Load Balancer
- Exportação de dados de log para o Amazon S3
- Instalação do agente do CloudWatch em uma instância do Amazon EC2
- Publicar métricas personalizadas
- Uso de painéis do Amazon CloudWatch
- Uso de métricas do Amazon CloudWatch
- Uso de canários (Amazon CloudWatch Synthetics)
- O que é o Amazon CloudWatch Logs?

Guias do usuário:

- Criação de uma trilha
- Monitoramento de métricas de memória e de disco para instâncias do Linux do Amazon EC2
- Uso do CloudWatch Logs com instâncias de contêiner
- Logs de fluxo da VPC
- O que é o Amazon DevOps Guru?
- O que é o AWS X-Ray?

Blogs relacionados:

- Depuração com o Amazon CloudWatch Synthetics e o AWS X-Ray

Exemplos e workshops relacionados:

- Laboratórios do AWS Well-Architected: excelência operacional: monitoramento de dependência

- A Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional
- Workshop de observabilidade

REL06-BP02 Definir e calcular 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.

O Amazon CloudWatch e o Amazon S3 funcionam como camadas primárias de agregação e armazenamento. Para alguns serviços, como o AWS Auto Scaling e o Elastic Load Balancing, métricas padrão são fornecidas para carga de CPU ou latência média de solicitação em um cluster ou uma instância. Para serviços de streaming, como o VPC Flow Logs e o AWS CloudTrail, dados de evento são encaminhados ao CloudWatch Logs, e você precisa definir e aplicar filtros de métricas para extraí-las dos dados do evento. Isso fornece dados de séries temporais, que podem servir como entradas para alarmes do CloudWatch que você define para acionar alertas.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- 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
  - Os filtros de métrica definem os termos e os padrões a serem procurados nos dados de log à medida que são enviados para o CloudWatch Logs. O CloudWatch Logs usa esses filtros para transformar dados de log em métricas numéricas do CloudWatch, que você pode representar graficamente ou para as quais pode definir um alarme.
    - Pesquisa e filtragem de dados de log
  - Use um terceiro confiável para agregar logs.
    - Siga as instruções do terceiro. A maioria dos produtos de terceiros integra-se ao CloudWatch e ao Amazon S3.
  - Alguns serviços da AWS podem publicar logs diretamente no Amazon S3. Se seu principal requisito de logs for o armazenamento no Amazon S3, você poderá facilmente fazer com que o serviço que produz os logs os envie diretamente ao Amazon S3 sem configurar uma infraestrutura adicional.
    - Envie logs diretamente ao Amazon S3

Recursos

Documentos relacionados:

- Consultas de exemplo do Amazon CloudWatch Logs Insights
- Depuração com o Amazon CloudWatch Synthetics e o AWS X-Ray
- Um workshop de observabilidade
- Pesquisa e filtragem de dados de log
- Envie logs diretamente ao Amazon S3
- A Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional

REL06-BP03 Enviar notificações (processamento e emissão de alarmes em tempo real)

As organizações que precisam estar a par de tudo, recebem notificações quando ocorrem eventos importantes.

Os alertas também podem ser enviados para tópicos do Amazon Simple Notification Service (Amazon SNS) e, em seguida, publicados para qualquer número de assinantes. Por exemplo, o Amazon SNS pode encaminhar alertas a um alias de e-mail para que a equipe técnica possa responder.

Antipadrões comuns:

- Configurar alarmes com um limite muito baixo, fazendo com que muitas notificações sejam enviadas.
- Não arquivar alarmes para exploração futura.

Benefícios do estabelecimento dessa prática recomendada: As notificações sobre eventos (mesmo aqueles que podem ser respondidos e resolvidos automaticamente) permitem que você tenha um registro dos eventos e, possivelmente, resolva-os de maneira diferente no futuro.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Realize o processamento e a emissão de alarmes em tempo real. As organizações que precisam estar a par de tudo, recebem notificações quando ocorrem eventos importantes

- Os painéis do Amazon CloudWatch são páginas iniciais personalizáveis no console do CloudWatch, que você pode usar para monitorar os recursos em uma única visualização, mesmo aqueles distribuídos por regiões diferentes.
  - Uso de painéis do Amazon CloudWatch
- Crie um alarme quando a métrica ultrapassar um limite.
  - Uso de alarmes do Amazon CloudWatch

Recursos

Documentos relacionados:

- Um workshop de observabilidade
- A Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional
- Uso de alarmes do Amazon CloudWatch
- Uso de painéis do Amazon CloudWatch
- Uso de métricas do Amazon CloudWatch

REL06-BP04 Automatizar 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.

Alertas podem acionar eventos do AWS Auto Scaling, para que os clusters possam reagir a alterações na demanda. Os alertas podem ser enviados ao Amazon Simple Queue Service (Amazon SQS), que pode servir como um ponto de integração para sistemas de tíquetes de terceiros. O AWS Lambda também pode assinar alertas, fornecendo aos usuários um modelo de tecnologia sem servidor assíncrono que reage a alterações dinamicamente. O AWS Config monitora e registra continuamente as configurações de recursos da AWS e pode acionar o AWS Systems Manager Automation para corrigir problemas.

O Amazon DevOps Guru pode monitorar automaticamente recursos de aplicações em busca de comportamento anômalo e entregar recomendações direcionadas para acelerar os tempos de identificação e correção de problemas.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Use o Amazon DevOps Guru para realizar ações automatizadas. O Amazon DevOps Guru pode monitorar automaticamente recursos de aplicações em busca de comportamento anômalo e entregar recomendações direcionadas para acelerar os tempos de identificação e correção de problemas.
  - O que é o Amazon DevOps Guru?
- Use o AWS Systems Manager para realizar ações automatizadas. O AWS Config monitora e registra continuamente as configurações de recursos da AWS e pode acionar o AWS Systems Manager Automation para corrigir problemas.
  - AWS Systems Manager Automation
    - Crie e use documentos do Systems Manager Automation. Eles definem as ações que o Systems Manager realiza nas suas instâncias gerenciadas e em outros recursos da AWS quando ocorre um processo de automação.
    - Trabalhar com documentos de automação (playbooks)
- O Amazon CloudWatch envia eventos de mudança de estado de alarme para o Amazon EventBridge. Crie regras do EventBridge para automatizar respostas.
  - Criar uma regra do EventBridge que seja acionada em um evento de um recurso da AWS
- Crie e execute um plano para automatizar respostas.
  - Faça o inventário de todos os seus procedimentos de resposta de alerta. Você deve planejar suas respostas de alerta antes de classificar as tarefas.
  - Faça o inventário de todas as tarefas com ações específicas que devem ser executadas. A maioria dessas ações está documentada nos runbooks. Você também deve ter playbooks para alertas de eventos inesperados.
  - Examine os runbooks e os playbooks de todas as ações automatizáveis. Em geral, se for possível definir uma ação, ela provavelmente poderá ser automatizada.
  - Classifique primeiro as atividades demoradas ou propensas a erros. É mais vantajoso remover as fontes de erros e reduzir o tempo de resolução.
  - Estabeleça um plano para concluir a automação. Mantenha um plano ativo para automatizar e atualizar a automação.
  - Examine os requisitos manuais para criar oportunidades de automação. Desafie seu processo manual para criar oportunidades de automatização.

Recursos

Documentos relacionados:

- AWS Systems Manager Automation
- Criar uma regra do EventBridge que seja acionada em um evento de um recurso da AWS
- Um workshop de observabilidade
- A Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional
- O que é o Amazon DevOps Guru?
- Trabalhar com documentos de automação (playbooks)

REL06-BP05 Análises

colete arquivos de log e históricos de métricas e analise-os para obter tendências mais abrangentes e informações sobre a carga de trabalho.

O Amazon CloudWatch Logs oferece suporte a uma linguagem de consulta simples, mas poderosa que você pode usar para analisar dados de log. O Amazon CloudWatch Logs também oferece suporte a assinaturas que permitem que os dados fluam perfeitamente ao Amazon S3, onde você pode usar o ou o Amazon Athena para consultar esses dados. Ele oferece suporte a consultas em uma grande variedade de formatos. Perceber Formatos de dados e SerDes compatíveis no guia do usuário do Amazon Athena para obter mais informações. Para análise de conjuntos enormes de arquivos de log, você pode executar um cluster do Amazon EMR para executar análises em escala de petabytes.

Existem várias ferramentas fornecidas por parceiros da AWS e por terceiros que permitem agregação, processamento, armazenamento e estudo analítico. Essas ferramentas incluem New Relic, Splunk, Loggly, Logstash, CloudHealth e Nagios. Porém, a geração fora dos registros do aplicativo e do sistema é única para cada provedor de nuvem e costuma ser única para cada serviço.

Uma parte do processo de monitoramento que costuma ser negligenciada é o gerenciamento de dados. Você precisa determinar os requisitos de retenção para monitorar os dados e então aplicar as políticas de ciclo de vida de acordo. O Amazon S3 oferece suporte ao gerenciamento de ciclo de vida no nível do bucket do S3. Esse gerenciamento de ciclo de vida pode ser aplicado de modo diferente a diferentes caminhos no bucket. Mais perto do fim do ciclo de vida, você pode fazer a transição dos dados ao Amazon S3 Glacier para armazenamento de longo prazo e posterior expiração após o fim do período de retenção. A classe de armazenamento S3 Intelligent-Tiering foi projetada para otimizar

custos movendo automaticamente dados para o nível de acesso mais econômico, sem impacto na performance ou sobrecarga operacional.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- O CloudWatch Logs Insights permite pesquisar e analisar dinamicamente seus dados de log no Amazon CloudWatch Logs.
  - Análise de dados de log com o CloudWatch Logs Insights
  - Consultas de exemplo do Amazon CloudWatch Logs Insights
- Use o Amazon CloudWatch Logs para enviar logs para o Amazon S3, onde você pode usar o Amazon Athena para consultar dados.
  - Como analiso meus logs de acesso ao servidor do Amazon S3 usando o Athena?
    - Crie uma política de ciclo de vida do S3 para o bucket de logs de acesso ao seu servidor. Configure a política de ciclo de vida para remover periodicamente os arquivos de log. Esse procedimento reduz a quantidade de dados que o Athena analisa em cada consulta.
      - Como faço para criar uma política de ciclo de vida de um bucket do S3?

Recursos

Documentos relacionados:

- Consultas de exemplo do Amazon CloudWatch Logs Insights
- Análise de dados de log com o CloudWatch Logs Insights
- Depuração com o Amazon CloudWatch Synthetics e o AWS X-Ray
- Como faço para criar uma política de ciclo de vida de um bucket do S3?
- Como analiso meus logs de acesso ao servidor do Amazon S3 usando o Athena?
- Um workshop de observabilidade
- A Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional

REL06-BP06 Realizar revisões regularmente

Revise frequentemente a implementação do monitoramento da workload e atualize-a com base em eventos e alterações significativos.

O monitoramento eficaz é orientado pelas principais métricas de negócios. Certifique-se de que essas métricas sejam acomodadas em sua carga de trabalho à medida que as prioridades de negócios mudam.

Auditar seu monitoramento ajuda a garantir que você saiba quando um aplicativo está atingindo as respectivas metas de disponibilidade. A análise da causa raiz requer a capacidade de descobrir o que aconteceu quando ocorreram falhas. A AWS fornece serviços que permitem acompanhar o estado dos seus serviços durante um incidente:

- Amazon CloudWatch Logs: você pode armazenar seus logs nesse serviço e inspecionar seu conteúdo.
- Amazon CloudWatch Logs Insights: é um serviço totalmente gerenciado que permite analisar logs massivos em segundos. Ele oferece consultas e visualizações rápidas e interativas.
- AWS Config: você pode ver qual infraestrutura da AWS estava em uso em diferentes momentos.
- AWS CloudTrail: você pode ver quais APIs da AWS foram invocadas, a que horas e por qual entidade principal.

Na AWS, realizamos uma reunião semanal para revisar a performance operacional e para compartilhar aprendizados entre as equipes. Como há tantas equipes na AWS, criamos A roda para escolher aleatoriamente uma carga de trabalho para revisão. Estabelecer um ritmo regular para análises de performance operacional e compartilhamento de conhecimento aprimora sua capacidade de obter uma performance superior de suas equipes operacionais.

Antipadrões comuns:

- Coletar apenas as métricas padrão.
- Definir uma estratégia de monitoramento e nunca revisá-la.
- Não analisar o monitoramento quando alterações importantes são implantadas.

Benefícios do estabelecimento dessa prática recomendada: A revisão regular do monitoramento permite a antecipação de possíveis problemas, em vez de reagir a notificações quando um problema previsto realmente ocorrer.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Crie vários painéis para a workload. Você deve ter um painel superior com as principais métricas de negócios e as métricas técnicas identificadas como as mais relevantes à integridade projetada da carga de trabalho conforme a variação do uso. Você também deve ter painéis para vários níveis e dependências da aplicação que podem ser inspecionados.
  - Uso de painéis do Amazon CloudWatch
- Programe e realize revisões regulares dos painéis da workload. Realize uma inspeção regular dos painéis. Você pode ter graus diferentes de profundidade para a inspeção.
  - Inspecione as tendências nas métricas. Compare os valores das métricas com os valores históricos para ver se há tendências que possam indicar algo que precise de investigação. Exemplos disso incluem: aumento da latência, diminuição da função principal de negócios e aumento das respostas a falhas.
  - Verifique se há exceções ou anomalias nas suas métricas. As médias ou os valores medianos podem mascarar as exceções e as anomalias. Examine os valores mais altos e mais baixos durante o período e investigue as causas das pontuações extremas. À medida que você continua a eliminar essas causas, a redução da definição de extremo permite melhorar cada vez mais a consistência da performance da workload.
  - Procure mudanças bruscas no comportamento. Uma mudança imediata na quantidade ou na direção de uma métrica pode indicar que houve uma alteração na aplicação ou fatores externos aos quais você talvez precise adicionar outras métricas para acompanhar.

Recursos

Documentos relacionados:

- Consultas de exemplo do Amazon CloudWatch Logs Insights
- Depuração com o Amazon CloudWatch Synthetics e o AWS X-Ray
- Um workshop de observabilidade
- A Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional
- Uso de painéis do Amazon CloudWatch

REL06-BP07 Monitorar 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 das aplicações e dos serviços subjacentes delas.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Monitore o rastreamento completo das solicitações por meio do seu sistema. O AWS X-Ray é um serviço que coleta dados sobre as solicitações atendidas por sua aplicação e fornece ferramentas que você pode usar para visualizar, filtrar e obter insights desses dados para identificar problemas e oportunidades de otimização. Para qualquer solicitação rastreada para a sua aplicação, você pode ver informações detalhadas sobre a solicitação e a resposta e também sobre as chamadas que a aplicação faz para recursos downstream da AWS, microsserviços, bancos de dados e APIs da Web.
  - O que é o AWS X-Ray?
  - Depuração com o Amazon CloudWatch Synthetics e o AWS X-Ray

Recursos

Documentos relacionados:

- Depuração com o Amazon CloudWatch Synthetics e o AWS X-Ray
- Um workshop de observabilidade
- A Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional
- Uso de canários (Amazon CloudWatch Synthetics)
- O que é o AWS X-Ray?

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.

Práticas recomendadas

- REL07-BP01 Usar a automação ao obter ou escalar recursos
- REL07-BP02 Obter recursos após a detecção de danos em uma workload
- REL07-BP03 Obter recursos após a detecção de que mais recursos são necessários para uma workload
- REL07-BP04 Fazer o teste de carga da sua workload

REL07-BP01 Usar a automação ao obter ou escalar recursos

Ao substituir recursos danificados ou escalar sua workload, 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 AWS SDKs para automatizar a escalabilidade.

Os serviços gerenciados pela AWS incluem o Amazon S3, o Amazon CloudFront, o AWS Auto Scaling, o AWS Lambda, o Amazon DynamoDB, o AWS Fargate e o Amazon Route 53.

O AWS Auto Scaling permite detectar e substituir instâncias danificadas. Ele também permite criar planos de escalabilidade para recursos, incluindo instâncias e frotas Spot do Amazon EC2 , tarefas do Amazon ECS tabelas e índices do Amazon DynamoDB e réplicas do Amazon Aurora .

Ao escalar instâncias do EC2, certifique-se de usar várias zonas de disponibilidade (de preferência, pelo menos três) e adicione ou remova capacidade para manter o equilíbrio entre essas zonas de disponibilidade. Tarefas do ECS ou pods do Kubernetes (ao usar o Amazon Elastic Kubernetes Service) também devem ser distribuídos em várias zonas de disponibilidade.

Ao usar o AWS Lambda, as instâncias são escaladas automaticamente. Sempre que uma notificação de evento é recebida para sua função, o AWS Lambda localiza rapidamente a capacidade livre dentro de sua frota de computação e executa seu código até a simultaneidade alocada. Você precisa se certificar de que a simultaneidade necessária esteja configurada no Lambda específico e no seu Service Quotas.

O Amazon S3 escala automaticamente para lidar com altas taxas de solicitação. Por exemplo, seu aplicativo pode atingir pelo menos 3.500 solicitações PUT/COPY/POST/DELETE ou 5.500 solicitações GET/HEAD por segundo por prefixo em um bucket. Não há limites para o número de prefixos em um bucket. Você pode aumentar a performance de leitura ou gravação paralelizando as leituras. Por exemplo, se você criar 10 prefixos em um bucket do Amazon S3 para paralelizar leituras, poderá escalar sua performance de leitura para 55 mil solicitações de leitura por segundo.

Configure e use o Amazon CloudFront ou uma rede de entrega de conteúdo (CDN) confiável. Uma CDN pode fornecer tempos mais rápidos de resposta ao usuário final e atender às solicitações de conteúdo do cache, reduzindo a necessidade de escalar a workload.

Antipadrões comuns:

- Implementar grupos de Auto Scaling para autorreparação, mas não implementar elasticidade.
- Usar a escalabilidade automática para responder a grandes aumentos no tráfego.
- Implantar aplicativos altamente com estado, eliminando a opção de elasticidade.

Benefícios do estabelecimento dessa prática recomendada: A automação elimina a possibilidade de erros manuais na implantação e no descomissionamento de recursos. A automação remove o risco de custos excedentes e de negação de serviço decorrentes da lentidão na resposta às necessidades de implantação ou de descomissionamento.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Configure e use o AWS Auto Scaling. Ele monitora seus aplicativos e ajusta automaticamente a capacidade para manter uma performance estável e previsível com o menor custo possível. Ao usar o AWS Auto Scaling, você pode configurar a escalabilidade da aplicação para vários recursos em diversos serviços.
  - O que é o AWS Auto Scaling?
    - Configure o Auto Scaling nas instâncias do Amazon EC2 e frotas spot, nas tarefas do Amazon ECS, nas tabelas e índices do Amazon DynamoDB, nas réplicas do Amazon Aurora e nos dispositivos do AWS Marketplace, conforme aplicável.
      - Gerenciamento da capacidade de throughput de modo automático com o DynamoDB Auto Scaling
        - Use as operações de API de serviço para especificar alarmes, políticas de escalabilidade e tempos de aquecimento e de resfriamento.
- Use o Elastic Load Balancing. Os load balancers podem distribuir a carga por caminho ou por conectividade de rede.
  - O que é o Elastic Load Balancing?
    - O Application Load Balancers pode distribuir a carga por caminho.
      - O que é um Application Load Balancer?

        - Configure um Application Load Balancer para distribuir o tráfego para diferentes workloads com base no caminho sob o nome de domínio.
        - É possível usar os Application Load Balancers para distribuir as cargas de maneira integrada ao AWS Auto Scaling para gerenciar a demanda.
            - Uso de um balanceador de carga com um grupo de Auto Scaling
    - Os Network Load Balancers podem distribuir a carga por conexão.
        - O que é um Network Load Balancer?
            - Configure um Network Load Balancer para distribuir o tráfego para cargas de trabalho diferentes por meio do TCP ou para ter um conjunto constante de endereços IP para a carga de trabalho.
            - É possível usar os Network Load Balancers para distribuir as cargas de maneira integrada ao AWS Auto Scaling para gerenciar a demanda.
- Use um provedor DNS altamente disponível. Nomes DNS permitem que os usuários insiram nomes, em vez de endereço IP, para acessar suas workloads e distribuem essas informações a um escopo definido, em geral, globalmente para usuários da workload.
    - Use o Amazon Route 53 ou um provedor DNS confiável.
        - O que é o Amazon Route 53?
    - Use o Route 53 para gerenciar as distribuições e os balanceadores de carga do CloudFront.
        - Determine os domínios e subdomínios que serão gerenciados.
        - Crie conjuntos de registros adequados com os registros ALIAS ou CNAME.
            - Trabalhando com registros
- Use a rede global da AWS para otimizar o caminho dos usuários às aplicações. O AWS Global Accelerator monitora continuamente a integridade dos endpoints da aplicação e redireciona o tráfego para endpoints íntegros em menos de 30 segundos.
    - O AWS Global Accelerator é um serviço que melhora a disponibilidade e a performance das aplicações com usuários locais ou globais. Ele fornece endereços IP estáticos que atuam como um ponto de entrada fixo para os endpoints da aplicação em uma ou várias Regiões da AWS, como os Application Load Balancers, os Network Load Balancers ou as instâncias do Amazon EC2.
        - O que é o AWS Global Accelerator?
- Configure e use o Amazon CloudFront ou uma rede de entrega de conteúdo (CDN) confiável. Uma rede de entrega de conteúdo pode fornecer tempos mais rápidos de resposta ao usuário final e

atender às solicitações de conteúdo que podem causar escalabilidade desnecessária das suas workloads.

- O que é o Amazon CloudFront?
  - Configure as distribuições do Amazon CloudFront para suas workloads ou use uma CDN de terceiros.
    - Você pode limitar o acesso às workloads para que elas sejam acessíveis somente pelo CloudFront usando os intervalos de IPs para o CloudFront nos seus grupos de segurança ou suas políticas de acesso de endpoint.

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudá-lo a criar soluções de computação automatizadas
- AWS Auto Scaling: como funcionam os planos de escalabilidade
- AWS Marketplace: produtos que podem ser usados com Auto Scaling
- Gerenciamento da capacidade de throughput de modo automático com o DynamoDB Auto Scaling
- Uso de um balanceador de carga com um grupo de Auto Scaling
- O que é o AWS Global Accelerator?
- O que é o Amazon EC2 Auto Scaling?
- O que é o AWS Auto Scaling?
- O que é o Amazon CloudFront?
- O que é o Amazon Route 53?
- O que é o Elastic Load Balancing?
- O que é um Network Load Balancer?
- O que é um Application Load Balancer?
- Trabalhando com registros

REL07-BP02 Obter recursos após a detecção de danos em uma workload

Escale recursos de modo reativo quando necessário, se a disponibilidade for afetada, para restaurar a disponibilidade da carga de trabalho.

Primeiro, você deve configurar as verificações de integridade e os critérios nessas verificações para indicar quando a disponibilidade é afetada pela falta de recursos. Em seguida, notifique o pessoal apropriado para escalar manualmente o recurso ou acione a automação para escalá-lo automaticamente.

A escala pode ser ajustada manualmente para sua workload. Por exemplo, é possível alterar o número de instâncias do EC2 em um grupo de Auto Scaling ou modificar o throughput de uma tabela do DynamoDB por meio do AWS Management Console ou do AWS CLI. No entanto, a automação deve ser usada sempre que possível (consulte Use a automação ao obter ou escalar recursos).

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Obtenha recursos após a detecção de danos em uma workload. Escale recursos de modo reativo quando necessário, se a disponibilidade for afetada, para restaurar a disponibilidade da carga de trabalho.
  - Use planos de escalabilidade, componente principal do AWS Auto Scaling, para configurar um conjunto de instruções para escalar seus recursos. Se você trabalha com o AWS CloudFormation ou adiciona tags aos recursos da AWS, poderá configurar planos de escalabilidade para diferentes conjuntos de recursos por aplicação. O AWS Auto Scaling fornece recomendações para estratégias de escalabilidade personalizadas para cada recurso. Depois que o plano de escalabilidade for criado, o AWS Auto Scaling combinará os métodos de escalabilidade dinâmica e preditiva para oferecer suporte à sua estratégia de escalabilidade.
    - AWS Auto Scaling: como funcionam os planos de escalabilidade
  - O Amazon EC2 Auto Scaling ajuda a garantir que você tenha o número correto de instâncias do Amazon EC2 disponíveis para processar a carga da aplicação. Você cria coleções de instâncias do EC2, chamadas de grupos de Auto Scaling. Você pode especificar o número mínimo de instâncias em cada grupo de Auto Scaling, e o Amazon EC2 Auto Scaling garante que o grupo nunca fique abaixo desse tamanho. Você pode especificar o número mínimo de instâncias em cada grupo de Auto Scaling, e o Amazon EC2 Auto Scaling garante que o grupo nunca fique abaixo desse tamanho.
    - O que é o Amazon EC2 Auto Scaling?
  - A escalabilidade automática do Amazon DynamoDB usa o serviço AWS Application Auto Scaling para ajustar dinamicamente a capacidade de throughput provisionado por você, em resposta aos padrões de tráfego reais. Isso permite que uma tabela ou um índice secundário

global aumente sua capacidade provisionada de leitura e gravação para sustentar aumentos repentinos no tráfego, sem controle de utilização.

- Gerenciamento da capacidade de throughput de modo automático com o DynamoDB Auto Scaling

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudá-lo a criar soluções de computação automatizadas
- AWS Auto Scaling: como funcionam os planos de escalabilidade
- AWS Marketplace: produtos que podem ser usados com Auto Scaling
- Gerenciamento da capacidade de throughput de modo automático com o DynamoDB Auto Scaling
- O que é o Amazon EC2 Auto Scaling?

REL07-BP03 Obter recursos após a detecção de que mais recursos são necessários para uma workload

Escale os recursos proativamente para atender à demanda e evitar impacto na disponibilidade.

Muitos serviços da AWS são escalados automaticamente para atender à demanda. Se estiver usando instâncias do Amazon EC2 ou clusters do Amazon ECS, você poderá configurar a escalabilidade automática deles para que ocorra com base nas métricas de uso que correspondam à demanda da workload. Para o Amazon EC2, a utilização média da CPU, a contagem de solicitações do load balancer ou a largura de banda da rede podem ser usadas para expandir (ou reduzir) instâncias do EC2. Para o Amazon ECS, a utilização média da CPU, a contagem de solicitações do balanceador de carga e a utilização da memória podem ser usadas para aumentar (ou reduzir) a escala horizontalmente de tarefas do ECS. Ao usar o Target Auto Scaling na AWS, o Autoscaler atua como um termostato doméstico, adicionando ou removendo recursos para manter o valor pretendido (por exemplo, 70% de utilização da CPU) que você especificar.

O AWS Auto Scaling também pode fazer o Auto Scaling preditivo, que usa machine learning para analisar a carga de trabalho histórica de cada recurso e prevê regularmente a carga futura para os próximos dois dias.

A Lei de Little ajuda a calcular quantas instâncias de computação (instâncias do EC2, funções simultâneas do Lambda etc.) são necessárias.

B = λW

L = número de instâncias (ou simultaneidade média no sistema)

λ = taxa média na qual as solicitações chegam (requisição por segundo)

W = tempo médio que cada solicitação gasta no sistema (s)

Por exemplo, a 100 rps, se cada solicitação demorar 0,5 segundos para ser processada, você precisará de 50 instâncias para acompanhar a demanda.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Obtenha recursos após a detecção de que mais recursos são necessários para uma workload. Escale os recursos proativamente para atender à demanda e evitar impacto na disponibilidade.
  - Calcule quantos recursos de computação serão necessários (simultaneidade de computação) para processar uma determinada taxa de solicitações.
    - Histórias sobre a Lei de Little
  - Quando você tiver um padrão histórico de uso, configure a escalabilidade programada para a escalabilidade automática do Amazon EC2.
    - Escalabilidade programada para o Amazon EC2 Auto Scaling
  - Use a escalabilidade preditiva da AWS.
    - Escalabilidade preditiva para o EC2 com Machine Learning

Recursos

Documentos relacionados:

- AWS Auto Scaling: como funcionam os planos de escalabilidade
- AWS Marketplace: produtos que podem ser usados com Auto Scaling
- Gerenciamento da capacidade de throughput de modo automático com o DynamoDB Auto Scaling
- Escalabilidade preditiva para o EC2 com Machine Learning
- Escalabilidade programada para o Amazon EC2 Auto Scaling
- Histórias sobre a Lei de Little
- O que é o Amazon EC2 Auto Scaling?

REL07-BP04 Fazer o teste de carga da sua workload

Adote uma metodologia de teste de carga para avaliar se a ação de escalabilidade atende aos requisitos da carga de trabalho.

É importante realizar testes de carga sustentada. Os testes de carga devem descobrir o ponto de interrupção e testar a performance da workload. A AWS facilita a configuração de ambientes de teste temporários que modelam a escala de sua workload 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.

Os testes de carga em produção também devem ser considerados como parte dos dias de jogos em que o sistema de produção é destacado, durante horas de menor utilização do cliente, com todo o pessoal disponível para interpretar os resultados e resolver os problemas que surgirem.

Antipadrões comuns:

- Executar testes de carga em implantações que não têm a mesma configuração da sua produção.
- Executar testes de carga apenas em componentes individuais da carga de trabalho, e não nela toda.
- Executar testes de carga com um subconjunto de solicitações, e não com um conjunto representativo de solicitações reais.
- Executar testes de carga para um pequeno fator de segurança acima da carga esperada.

Benefícios do estabelecimento dessa prática recomendada: Você sabe quais componentes em sua arquitetura falham sob carga e pode identificar as métricas que devem ser observadas para indicar que você está se aproximando dessa carga a tempo de resolver o problema, evitando o impacto dessa falha.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Realize testes de carga para identificar qual aspecto da workload indica que é necessário adicionar ou remover capacidade. Os testes de carga devem ter tráfego representativo semelhante ao que você recebe na produção. Aumente a carga enquanto observa as métricas que você preparou para determinar aquelas que indicam quando é necessário adicionar ou remover recursos.

- Teste de carga distribuída na AWS: simular milhares de usuários conectados
  - Identifique a combinação de solicitações. Você pode ter diversas combinações de solicitações, portanto, deve examinar vários períodos ao identificar a combinação de tráfego.
  - Implemente um direcionador de carga. Você pode usar um código personalizado, um código aberto ou um software comercial para implementar um direcionador de carga.
  - Faça o teste de carga inicialmente com uma pequena capacidade. Você vê alguns efeitos imediatos ao direcionar a carga para uma capacidade menor, possivelmente tão pequena quanto uma instância ou um contêiner.
  - Faça o teste de carga com uma capacidade maior. Os efeitos serão diferentes em uma carga distribuída, portanto, você deve testar o mais próximo possível de um ambiente de produto.

Recursos

Documentos relacionados:

- Teste de carga distribuída na AWS: simular milhares de usuários conectados

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.

Práticas recomendadas

- REL08-BP01 Usar runbooks para atividades padrão, como implantação
- REL08-BP02 Integrar testes funcionais como parte da sua implantação
- REL08-BP03 Integrar testes de resiliência como parte da sua implantação
- REL08-BP04 Implantar usando uma infraestrutura imutável
- REL08-BP05 Implantar alterações com automação

REL08-BP01 Usar runbooks para atividades padrão, como implantação

Os runbooks são os procedimentos predefinidos para alcançar um resultado específico. Use-os para executar atividades padrão, sejam elas feitas manualmente ou automaticamente. Os

exemplos incluem a implantação de uma workload, a aplicação de patches a ela ou a realização de modificações de DNS.

Por exemplo, coloque processos em vigor para garantir a segurança de reversão durante implantações. Garantir que você possa reverter uma implantação sem qualquer interrupção para seus clientes é essencial para tornar um serviço confiável.

Para procedimentos de runbooks, comece com um processo manual efetivo válido, implemente-o em código e acione-o para ser executado automaticamente quando adequado.

Mesmo para cargas de trabalho sofisticadas altamente automatizadas, os runbooks ainda são úteis para organizar dias de jogos ou atender a requisitos rigorosos de relatórios e auditoria.

Observe que playbooks são usados em resposta a incidentes específicos, e runbooks são usados para alcançar resultados específicos. Muitas vezes, os runbooks são para atividades de rotina, enquanto os playbooks são usados para responder a eventos que não são rotineiras.

Antipadrões comuns:

- Executar alterações não planejadas na configuração em produção.
- Ignorar as etapas do seu plano para agilizar a implantação, resultando em falha na implantação.
- Fazer alterações sem testar a inversão delas.

Benefícios do estabelecimento desta prática recomendada: O planejamento eficaz da alteração aumenta sua capacidade de executá-la com êxito, porque você está ciente de todos os sistemas afetados. A validação da alteração em ambientes de teste aumenta sua confiança.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Documente os procedimentos em runbooks para permitir respostas consistentes e rápidas a eventos bem conhecidos.
  - AWS Well-Architected Framework: conceitos: runbook
- Use o princípio de infraestrutura como código para definir sua infraestrutura. Ao usar o AWS CloudFormation (ou um terceiro confiável) para definir a infraestrutura, você poderá usar o software de controle de versão para controlar as versões e acompanhar as alterações.
  - Use o AWS CloudFormation (ou um provedor confiável de terceiros) para definir sua infraestrutura.

    - O que é o AWS CloudFormation?
- Use bons princípios de design de software para criar modelos exclusivos e desacoplados.
    - Determine as permissões, os modelos e as partes responsáveis pela implementação.
        - Controle de acesso com o AWS Identity and Access Management
    - Use o controle de origem, como o AWS CodeCommit ou uma ferramenta confiável de terceiros, para controle de versão.
        - O que é o AWS CodeCommit?

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudá-lo a criar soluções de implantação automatizada
- AWS Marketplace: produtos que podem ser usados para automatizar suas implantações
- AWS Well-Architected Framework: conceitos: runbook
- O que é o AWS CloudFormation?
- O que é o AWS CodeCommit?

Exemplos relacionados:

- Automatização de operações com playbooks e runbooks

REL08-BP02 Integrar 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.

Esses testes são executados em um ambiente de pré-produção, que é preparado antes da produção no pipeline. Idealmente, isso é feito como parte de um pipeline de implantação.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- 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.

- Invoque o AWS CodeBuild durante a “ação de teste” dos pipelines de lançamento de software baseados no AWS CodePipeline. Esse recurso permite que você execute facilmente uma variedade de testes no código, como testes de unidade, análises de código estático e testes de integração.
  - O AWS CodePipeline adiciona compatibilidade para testes de unidade e de integração personalizada com o AWS CodeBuild
- Use as soluções do AWS Marketplace para executar testes automatizados como parte do pipeline de entrega de software.
  - Automação de teste de software

Recursos

Documentos relacionados:

- O AWS CodePipeline adiciona compatibilidade para testes de unidade e de integração personalizada com o AWS CodeBuild
- Automação de teste de software
- O que é o AWS CodePipeline?

REL08-BP03 Integrar testes de resiliência como parte da sua implantação

Os testes de resiliência (usando os princípios da engenharia do caos) são executados como parte do pipeline de implantação automatizado em um ambiente de pré-produção.

Esses testes são preparados e executados no pipeline em um ambiente de pré-produção. Eles também devem ser executados em produção como parte de dias de jogo.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Integre testes de resiliência como parte da sua implantação. Use a engenharia do caos, a disciplina de experimentar em uma workload, para gerar confiança na capacidade da workload de resistir a condições conturbadas na produção.
  - Os testes de resiliência injetam falhas ou degradação de recursos para avaliar se a workload responde com a resiliência projetada.
    - Laboratório do Well-Architected: nível 300: testes de resiliência do EC2 RDS e do S3

- Esses testes podem ser executados regularmente em ambientes de pré-produção nos pipelines de implantação automatizados.
- Eles também devem ser executados em produção, como parte dos dias de jogo programados.
- Ao adotar os princípios da engenharia do caos, proponha hipóteses de como a carga de trabalho será executada sob várias condições adversas e, em seguida, teste essas hipóteses por meio dos testes de resiliência.
  - Princípios da engenharia do caos

Recursos

Documentos relacionados:

- Princípios da engenharia do caos
- O que é o AWS Fault Injection Simulator?

Exemplos relacionados:

- Laboratório do Well-Architected: nível 300: testes de resiliência do EC2 RDS e do S3

REL08-BP04 Implantar usando uma infraestrutura imutável

A infraestrutura imutável é um modelo que não requer atualizações, patches de segurança ou alterações de configuração nas workloads de produção. Quando uma alteração é necessária, a arquitetura é criada em uma nova infraestrutura e implantada na produção.

A implementação mais comum do paradigma de infraestrutura imutável é o servidor imutável. Isso significa que, se um servidor precisar de uma atualização ou uma correção, novos servidores serão implantados em vez de atualizar os já em uso. Portanto, em vez de fazer login no servidor via SSH e atualizar a versão do software, cada alteração no aplicativo começa com um push de software para o repositório de código, por exemplo, git push. Como as alterações não são permitidas na infraestrutura imutável, você pode ter certeza sobre o estado do sistema implantado. As infraestruturas imutáveis são inerentemente mais consistentes, confiáveis e previsíveis, e simplificam muitos aspectos do desenvolvimento e operações de software.

Use uma implantação de canário ou azul/verde ao implantar aplicativos em infraestruturas imutáveis.

Implantação canário é a prática de direcionar um pequeno número de seus clientes para a nova versão, geralmente em execução em uma única instância de serviço (o canário). Em seguida, você examina profundamente todas as alterações de comportamento ou erros gerados. Você poderá remover o tráfego da implantação canário se encontrar problemas críticos e enviar os usuários de volta para a versão anterior. Se a implantação for bem-sucedida, você poderá continuar implantando a uma velocidade desejada enquanto monitora as alterações em busca de erros até a implantação estar concluída. O AWS CodeDeploy pode ser configurado com uma configuração de implantação que permitirá uma implantação canário.

A implantação azul/verde é semelhante à implantação canário, com a diferença que um conjunto completo do aplicativo é implantado em paralelo. Você alterna as implantações entre as duas pilhas (azul e verde). Novamente, é possível enviar o tráfego para a nova versão e voltar para a versão antiga se houver problemas na implantação. Normalmente, todo o tráfego é alternado de uma só vez. No entanto, você também pode usar frações do tráfego para cada versão para aumentar a adoção da nova versão usando os recursos de roteamento de DNS ponderado do Amazon Route 53. O AWS CodeDeploy e o AWS Elastic Beanstalk podem ser definidos com uma configuração de implantação que permitirá uma implantação azul/verde.

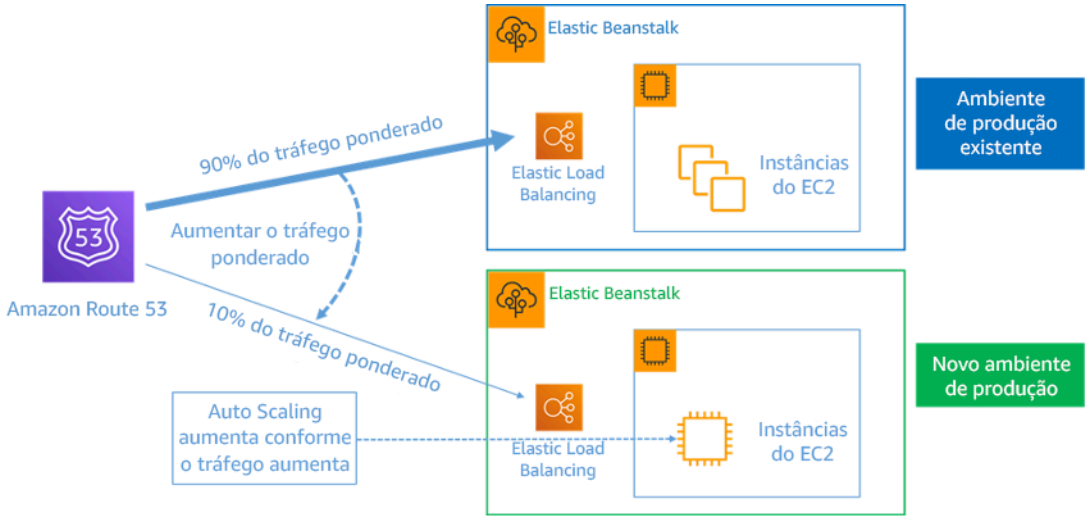


Figura 8: Implantação azul/verde com o AWS Elastic Beanstalk e o Amazon Route 53

Benefícios da infraestrutura imutável:

- Redução em desvios de configuração: ao substituir frequentemente os servidores de uma configuração básica, conhecida e controlada por versão, a infraestrutura é redefinida para um estado conhecido, evitando desvios de configuração.
- Implantações simplificadas: as implantações são simplificadas porque não precisam oferecer suporte a atualizações. As atualizações são apenas novas implantações.

- Implantações atômicas confiáveis: as implantações são concluídas com êxito ou nada muda. Ele dá mais confiança no processo de implantação.
- Implantações mais seguras com processos rápidos de reversão e recuperação: as implantações são mais seguras, pois a versão de trabalho anterior não é alterada. Você pode reverter para ele se forem detectados erros.
- Ambientes consistentes de teste e depuração: como todos os servidores usam a mesma imagem, não há diferenças entre ambientes. Uma compilação é implantada em vários ambientes. Ele também evita ambientes inconsistentes e simplifica o teste e a depuração.
- Maior escalabilidade: como os servidores usam uma imagem base, são consistentes e podem ser repetidos, a escalabilidade automática é trivial.
- Cadeia de ferramentas simplificada: a cadeia de ferramentas é simplificada, pois você pode se livrar das ferramentas de gerenciamento de configuração gerenciando atualizações de software de produção. Não há ferramentas ou agentes adicionais instalados nos servidores. As alterações são feitas na imagem base, testadas e implementadas.
- Maior segurança: ao negar todas as alterações nos servidores, você pode desabilitar o SSH nas instâncias e remover chaves. Isso reduz o vetor de ataque, melhorando a postura de segurança da sua organização.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Faça a implantação com uma infraestrutura imutável. A infraestrutura imutável é um modelo no qual não ocorrem atualizações, patches de segurança ou alterações de configuração no local em sistemas de produção. Se for necessária uma alteração, outra versão da arquitetura será criada e implantada na produção.
  - Visão geral de uma implantação azul/verde
  - Implantação gradual de aplicativos sem servidor
  - Infraestrutura imutável: confiabilidade, consistência e confiança por meio da imutabilidade
  - CanaryRelease

Recursos

Documentos relacionados:

- CanaryRelease

- Implantação gradual de aplicativos sem servidor
- Infraestrutura imutável: confiabilidade, consistência e confiança por meio da imutabilidade
- Visão geral de uma implantação azul/verde
- A Amazon Builders' Library: garanta a segurança da reversão durante implantações

REL08-BP05 Implantar alterações com automação

As implantações e a aplicação de patches são automatizadas para eliminar o impacto negativo.

As alterações nos sistemas de produção são uma das maiores áreas de risco para muitas organizações. Consideramos as implantações um problema de primeira classe a ser resolvido junto com os problemas de negócio que o software aborda. Atualmente, isso significa usar a automação nas operações sempre que for viável, incluindo testar e implantar alterações, adicionar ou remover capacidade e migrar dados. O AWS CodePipeline permite gerenciar as etapas necessárias para liberar a sua carga de trabalho. Isso inclui um estado de implantação usando o AWS CodeDeploy para automatizar a implantação do código do aplicativo em instâncias do Amazon EC2, instâncias on-premises, funções do Lambda sem servidor ou serviços do Amazon ECS.

> Recomendação
>
> Embora a sabedoria convencional sugira que você mantenha humanos no ciclo para os procedimentos operacionais mais difíceis, sugerimos automatizar esses procedimentos exatamente por isso.

Antipadrões comuns:

- Executar as alterações manualmente.
- Ignorar as etapas da sua automação por meio de fluxos de trabalho de emergência.
- Não seguir seus planos.

Benefícios do estabelecimento desta prática recomendada: Ao usar a automação para implantar todas as alterações, você elimina as chances de introduzir erros humanos e permite que sejam feitos testes antes de alterar a produção para garantir que seus planos sejam conduzidos.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Automatize seu pipeline de implantação. Os pipelines de implantação permitem invocar testes automatizados e detecção de anomalias. Além disso, eles interrompem o pipeline em uma determinada etapa antes da implantação em produção ou revertem automaticamente uma alteração.
  - A Amazon Builders' Library: garanta a segurança da reversão durante implantações
  - A Amazon Builders' Library: acelere a entrega contínua
    - Use o AWS CodePipeline (ou um produto de terceiros confiável) para definir e executar seus pipelines.
      - Configure o pipeline para ser iniciado quando uma alteração for confirmada no repositório do seu código.
        - O que é o AWS CodePipeline?
      - Use o Amazon Simple Notification Service (Amazon SNS) e o Amazon Simple Email Service (Amazon SES) para enviar notificações sobre problemas no pipeline ou integrar-se com uma ferramenta de bate-papo da equipe, como o Amazon Chime.
        - O que é o Amazon Simple Notification Service?
        - O que é o Amazon SES?
        - O que é o Amazon Chime?
        - Automatize mensagens de chat com webhooks.

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudá-lo a criar soluções de implantação automatizada
- AWS Marketplace: produtos que podem ser usados para automatizar suas implantações
- Automatize mensagens de chat com webhooks.
- A Amazon Builders' Library: garanta a segurança da reversão durante implantações
- A Amazon Builders' Library: acelere a entrega contínua
- O que é o AWS CodePipeline?
- O que é o CodeDeploy?
- AWS Systems Manager Patch Manager
- O que é o Amazon SES?

- O que é o Amazon Simple Notification Service?

Vídeos relacionados:

- Conferência da AWS 2019: CI/CD na AWS

Gerenciamento de falhas

Perguntas

- REL 9 Como você faz backup dos dados?
- REL 10 Como usar o isolamento de falhas para proteger sua carga de trabalho?
- REL 11 Como você projeta sua carga de trabalho para resistir a falhas de componentes?
- REL 12 Como testar a confiabilidade?
- REL 13 Como você planeja a recuperação de desastres (DR)?

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).

Práticas recomendadas

- REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes
- REL09-BP02 Proteger e criptografar backups
- REL09-BP03 Realizar o backup de dados automaticamente
- REL09-BP04 Realizar a recuperação periódica dos dados para verificar a integridade e os processos de backup

REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes

Todos os armazenamentos de dados da AWS oferecem recursos de backup. Serviços como o Amazon RDS e o Amazon DynamoDB oferecem suporte adicional ao backup automatizado que

permite a recuperação a um ponto anterior no tempo (PITR), ajudando você a restaurar um backup a qualquer momento até cinco minutos ou menos, antes da hora atual. Muitos serviços da AWS oferecem a capacidade de copiar backups para outra Região da AWS. O AWS Backup é uma ferramenta que fornece a capacidade de centralizar e automatizar a proteção de dados em todos os serviços da AWS.

É possível usar o Amazon S3 como destino de backup para fontes de dados autogerenciadas e gerenciadas pela AWS. Os serviços da AWS como o Amazon EBS, o Amazon RDS e o Amazon DynamoDB, têm recursos integrados para criar backups. É possível também usar um software de backup de terceiros.

É possível fazer backup de dados on-premises na Nuvem AWS usando AWS Storage Gateway ou AWS DataSync. É possível usar os buckets do Amazon S3 para armazenar estes dados na AWS. O Amazon S3 oferece vários níveis de armazenamento, como Amazon S3 Glacier ou S3 Glacier Deep Archive , para reduzir custos de armazenamento de dados.

Você pode atender às necessidades de recuperação de dados reproduzindo os dados de outras fontes. Por exemplo: os nós de réplicas do Amazon ElastiCache ou as réplicas de leitura do RDS podem ser usados para reproduzir dados caso o primário seja perdido. Nos casos em que fontes como esta podem ser usadas para atender aos seus objetivo de tempo de recuperação (RTO) e objetivo de ponto de recuperação (RPO), pode ser que você não precise fazer backup. Outro exemplo: se estiver trabalhando com Amazon EMR, poderá não ser necessário fazer backup do seu armazenamento de dados HDFS, desde que você possa reproduzir os dados no EMR do S3.

Ao selecionar uma estratégia de backup, considere o tempo necessário para recuperar os dados. Ele depende do tipo de backup (no caso de uma estratégia de backup) ou da complexidade do mecanismo de reprodução de dados. O tempo deve estar dentro do RTO para a workload.

Resultado desejado:

As fontes de dados foram identificadas e classificadas com base na criticidade. Em seguida, estabeleça uma estratégia de recuperação de dados com base no RPO. A estratégia envolve fazer o backup dessas fontes de dados ou a capacidade de reproduzir dados de outras fontes. Em caso de perda de dados, a estratégia implementada permite a recuperação ou reprodução de dados dentro do RPO e RTO definidos.

Fase de maturidade da nuvem: Foundational

Antipadrões comuns:

- Não estar ciente de todas as fontes de dados para a workload e sua criticidade.
- Não fazer backups de fontes de dados essenciais.
- Fazer backups apenas de algumas fontes de dados sem usar a criticidade como critério.
- Não ter um RPO definido ou a frequência de backup não atender ao RPO.
- Não avaliar a necessidade de um backup ou se os dados podem ser reproduzidos de outras fontes.

Benefícios do estabelecimento dessa prática recomendada: Identificar os locais onde os backups são necessários, implementar um mecanismo para criar backups ou poder reproduzir os dados de uma fonte externa melhora a capacidade de restaurar e recuperar dados durante uma interrupção.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

Compreenda e use os recursos de backup dos serviços e recursos da AWS usados pela workload. A maioria dos serviços da AWS oferece recursos para fazer backup dos dados da workload.

Etapas da implementação:

1. Identifique todas as fontes de dados para a workload. Os dados podem ser armazenados em vários recursos, como relacional, volumes, sistemas de arquivos, sistemas de registro em logse aos armazenamento de objeto. Consulte o Recursos seção para encontrar Documentos relacionados a diferentes serviços da AWS onde os dados são armazenados e a capacidade de fazer backup que eles fornecem.
2. Classifique as fontes de dados com base na criticidade. Diferentes conjuntos de dados terão diferentes níveis de criticidade para uma workload e, portanto, diferentes requisitos de resiliência. Por exemplo, alguns dados podem ser críticos e exigir um RPO próximo de zero, enquanto outros dados podem ser menos críticos e tolerar um RPO mais alto e a perda de alguns dados. Da mesma forma, diferentes conjuntos de dados também podem ter diferentes requisitos de RTO.
3. Use a AWS ou os serviços de terceiros para criar backups dos dados. AWS Backup é um serviço gerenciado que permite criar backups de várias fontes de dados na AWS. A maioria desses serviços também possui recursos nativos para criar backups. O AWS Marketplace tem muitas soluções que também fornecem esses recursos. Consulte o Recursos listados abaixo para obter informações sobre como criar backups de dados de vários serviços da AWS.
4. Para dados sem backup, estabeleça um mecanismo de reprodução de dados.. Você pode optar por não fazer backup dos dados que podem ser reproduzidos de outras fontes por vários motivos.

Ás vezes, pode ser mais barato reproduzir dados de fontes se necessário, em vez de criar um backup, pois pode haver um custo associado ao armazenamento de backups. Outro exemplo é quando a restauração de um backup demora mais do que a reprodução dos dados das fontes, resultando em uma violação no RTO. Nestas situações, considere concessões e estabeleça um processo bem definido de como os dados podem ser reproduzidos dessas fontes quando a recuperação de dados for necessária. Por exemplo, se você carregou dados do Amazon S3 para um data warehouse (como o Amazon Redshift) ou para um cluster MapReduce (como o Amazon EMR) para analisá-los, esse é um exemplo de dados que podem ser reproduzidos de outras fontes. Desde que os resultados dessas análises sejam armazenados em algum lugar ou reproduzíveis, você não sofreria uma perda de dados devido a uma falha no data warehouse ou no cluster do MapReduce. Outros exemplos que podem ser reproduzidos de origens incluem caches (como o Amazon ElastiCache) ou réplicas de leitura do RDS.

5. Estabeleça um ritmo para fazer backup de dados. A criação de backups de fontes de dados é um processo periódico, e a frequência deve depender do RPO.

Nível de esforço para o plano de implementação: Moderado

Recursos

Práticas recomendadas relacionadas:

REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados

REL13-BP02 Usar estratégias de recuperação definidas para atender aos objetivos de recuperação

Documentos relacionados:

- O que é o AWS Backup?
- O que é o AWS DataSync?
- O que é o Gateway de Volumes?
- Parceiro do APN: parceiros que podem ajudar com o backup
- AWS Marketplace: produtos que podem ser usados para backup
- Snapshots do Amazon EBS
- Fazer backup do Amazon EFS
- Fazer backup do Amazon FSx for Windows File Server
- Backup e restauração para o ElastiCache for Redis

- Criar um snapshot do cluster de banco de dados no Neptune
- Criar um snapshot do banco de dados
- Criar uma regra do EventBridge que é acionada de acordo com uma programação
- Replicação entre regiões com o Amazon S3
- EFS-to-EFS AWS Backup
- Exportação de dados de log para o Amazon S3
- Gerenciamento do ciclo de vida de objetos
- Backup e restauração sob demanda para o DynamoDB
- Recuperação a um ponto anterior no tempo para o DynamoDB
- Como trabalhar com snapshots de índice do Amazon OpenSearch Service

Vídeos relacionados:

- AWS re:Invent 2021 - Backup, disaster recovery, and ransomware protection with AWS
- AWS Backup Demo: Cross-Account and Cross-Region Backup
- AWS re:Invent 2019: Deep dive on AWS Backup, ft. Rackspace (STG341)

Exemplos relacionados:

- Laboratório do Well-Architected: implementação da replicação bidirecional entre regiões (CRR) para o Amazon S3
- Laboratório do Well-Architected: teste de backup e restauração de dados
- Laboratório do Well-Architected: backup e restauração com failback para workload do Analytics
- Laboratório do Well-Architected: recuperação de desastres: backup e restauração

REL09-BP02 Proteger e criptografar backups

Controle e detecte o acesso a backups usando autenticação e autorização, como o AWS IAM. Use a criptografia para prevenir e detectar se a integridade dos dados de backups está comprometida.

O Amazon S3 oferece suporte a vários métodos de criptografia de dados ociosos. Ao usar a criptografia no lado do servidor, o Amazon S3 aceita seus objetos como dados não criptografados e, em seguida, criptografa-os ao armazená-los. Ao usar a criptografia do lado do cliente, a aplicação

da workload é responsável por criptografar os dados antes de serem enviados ao Amazon S3. Ambos os métodos permitem que você use o AWS Key Management Service (AWS KMS) para criar e armazenar a chave de dados, ou você pode fornecer sua própria chave, pela qual você é responsável. Usando o AWS KMS, você pode definir políticas usando o IAM sobre quem pode e não pode acessar suas chaves de dados e dados descriptografados.

Para o Amazon RDS, se você tiver optado por criptografar seus bancos de dados, seus backups também serão criptografados. Os backups do DynamoDB sempre são criptografados.

Antipadrões comuns:

- Ter o mesmo acesso aos backups e à automação de restauração que os dados.
- Não criptografar seus backups.

Benefícios do estabelecimento dessa prática recomendada: A proteção dos backups impede a violação dos dados, e a criptografia dos dados impede o acesso a eles se forem expostos por engano.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Use a criptografia em cada um dos seus armazenamentos de dados. Se os dados de origem forem criptografados, o backup também será.
  - Habilite a criptografia no RDS. Você pode configurar a criptografia em repouso usando o AWS Key Management Service ao criar uma instância do RDS.
    - Criptografia de recursos do Amazon RDS
  - Habilite a criptografia nos volumes do EBS. Você pode configurar a criptografia padrão ou especificar uma chave exclusiva após a criação do volume.
    - Criptografia do Amazon EBS
  - Use a criptografia necessário do Amazon DynamoDB. O DynamoDB criptografa todos os dados em repouso. Você pode usar uma chave AWS KMS de propriedade da AWS ou uma chave KMS gerenciada pela AWS, especificando uma chave armazenada na sua conta.
    - Criptografia em repouso do DynamoDB
    - Gerenciamento de tabelas criptografadas
  - Criptografe seus dados armazenados no Amazon EFS. Configure a criptografia ao criar seu sistema de arquivos.

    - Criptografia de dados e metadados no EFS
  - Configure a criptografia nas regiões de origem e de destino. Você pode configurar a criptografia em repouso no Amazon S3 usando as chaves armazenadas no KMS, mas as chaves são específicas da região. Você pode especificar as chaves de destino ao configurar a replicação.
    - Configuração adicional de CRR: replicação de objetos criados com a criptografia do lado do servidor (SSE) usando as chaves de criptografia armazenadas no AWS KMS
- Implemente permissões de privilégio mínimo para acessar seus backups. Siga as melhores práticas para limitar o acesso aos backups, aos snapshots e às réplicas de acordo com as melhores práticas de segurança.
  - Pilar Segurança: AWS Well-Architected

Recursos

Documentos relacionados:

- AWS Marketplace: produtos que podem ser usados para backup
- Criptografia do Amazon EBS
- Amazon S3: proteção de dados usando criptografia
- Configuração adicional de CRR: replicação de objetos criados com a criptografia do lado do servidor (SSE) usando as chaves de criptografia armazenadas no AWS KMS
- Criptografia em repouso do DynamoDB
- Criptografia de recursos do Amazon RDS
- Criptografia de dados e metadados no EFS
- Criptografia para backups na AWS
- Gerenciamento de tabelas criptografadas
- Pilar Segurança: AWS Well-Architected

Exemplos relacionados:

- Laboratório do Well-Architected: implementação da replicação bidirecional entre regiões (CRR) para o Amazon S3

REL09-BP03 Realizar o backup de dados automaticamente

Configure os backups para serem feitos automaticamente com base em uma programação periódica informadas pelo objetivo de ponto de recuperação (RPO) ou de acordo com alterações no conjunto de dados. É necessário fazer backup de conjuntos de dados críticos com requisitos de baixa perda de dados automaticamente com frequência, enquanto o backup de dados menos críticos, em que alguma perda é aceitável, pode ser feito com menos frequência.

É possível usar o AWS Backup para criar backups de dados automatizados de várias fontes de dados da AWS. É possível fazer backup das instâncias do Amazon RDS quase continuamente a cada cinco minutos e dos objetos do Amazon S3 a cada quinze minutos, proporcionando recuperação a um ponto anterior no tempo (PITR) para um momento específico no histórico de backup. Para outras fontes de dados da AWS, como volumes do Amazon EBS, tabelas do Amazon DynamoDB ou sistemas de arquivos do Amazon FSx, o AWS Backup pode executar backup automatizado de hora em hora. Esses serviços também oferecem recursos de backup nativos. Os serviços da AWS que oferecem backup automatizado com recuperação a um ponto anterior no tempo incluem Amazon DynamoDB, Amazon RDSe aos Amazon Keyspaces (para Apache Cassandra) . Eles podem ser restaurados a um momento específico no histórico do backup. A maioria dos outros serviços de armazenamento de dados da AWS oferece a capacidade de programar backups periódicos, até de hora em hora.

O Amazon RDS e o Amazon DynamoDB permitem o backup contínuo com recuperação a um ponto anterior no tempo. O versionamento do Amazon S3, uma vez habilitado, é automático. Amazon Data Lifecycle Manager pode ser usado para automatizar a criação, cópia e exclusão de snapshots do Amazon EBS. Ele também pode automatizar a criação, cópia, suspensão e cancelamento de imagens de máquina da Amazon (AMIs) suportadas pelo Amazon EBS e seus snapshots básicos do Amazon EBS.

Para obter uma visão centralizada da automação e do histórico de backups, o AWS Backup oferece uma solução de backup totalmente gerenciada e baseada em políticas. Ele centraliza e automatiza o backup de dados em vários serviços da AWS, na nuvem e on-premises, usando o AWS Storage Gateway.

Além do versionamento, o Amazon S3 oferece replicação. Todo o bucket do S3 pode ser replicado automaticamente para outro bucket na mesma Região da AWS ou em uma diferente.

Resultado desejado:

Um processo automatizado que cria backups de fontes de dados em um ritmo estabelecido.

Antipadrões comuns:

- Fazer backups manualmente.
- Usar recursos que têm o recurso de backup, mas não incluir o backup em sua automação.

Benefícios do estabelecimento desta prática recomendada: A automação de backups garante que eles sejam feitos regularmente com base no RPO e alerta você caso isso não ocorra.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

1. Identifique as fontes de dados que estão sendo copiados manualmente. Consulte REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes para orientações sobre isso.
2. Determine o RPO para a workload. Consulte REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados para orientações sobre isso.
3. Use uma solução de backup automatizada ou serviço gerenciado. O AWS Backup é um serviço totalmente gerenciado que facilita a centralização e automatização da proteção de dados em todos os serviços da AWS, na nuvem ou on-premises. Os planos de backup são um recurso do AWS Backup que permite a criação de regras que definem os recursos para backup e a frequência com que eles devem ser criados. A frequência deve ser informada pelo RPO estabelecido na etapa 2. Este laboratório do WA fornece orientação prática sobre como criar backups automatizados usando o AWS Backup. A maioria dos serviços da AWS que armazenam dados oferecem recursos de backup nativos. Por exemplo, o RDS pode ser aproveitado para backups automatizados com recuperação a um ponto anterior no tempo (PITR).
4. Para fontes de dados não suportadas por uma solução de backup automatizado ou serviço gerenciado, como fontes de dados ou filas de mensagens on-premises, considere usar uma solução confiável de terceiros para criar backups automatizados. Como alternativa, você pode criar automação para fazer isso usando a AWS CLI ou os SDKs. Você pode usar o AWS Lambda Functions ou o AWS Step Functions para definir a lógica envolvida na criação de um backup de dados e o Amazon EventBridge para executá-la em uma frequência baseada no RPO (conforme estabelecido na etapa 2).

Nível de esforço para o plano de implementação: Baixo

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar com o backup
- AWS Marketplace: produtos que podem ser usados para backup
- Criar uma regra do EventBridge que é acionada de acordo com uma programação
- O que é o AWS Backup?
- O que é o AWS Step Functions?

Vídeos relacionados:

- AWS re:Invent 2019: Deep dive on AWS Backup, ft. Rackspace (STG341)

Exemplos relacionados:

- Laboratório do Well-Architected: teste de backup e restauração de dados

REL09-BP04 Realizar 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.

Ao usar a AWS, você pode criar um ambiente de teste e restaurar seus backups para avaliar os recursos de RTO e RPO e executar testes de conteúdo e integridade dos dados.

Além disso, o Amazon RDS e o Amazon DynamoDB permitem a recuperação point-in-time (PITR). Ao usar o backup contínuo, você pode restaurar o conjunto de dados para o estado em que estava em uma data e hora especificadas.

Resultado desejado: Os dados de backups são recuperados periodicamente usando mecanismos bem definidos para garantir que a recuperação seja possível dentro do objetivo de tempo de recuperação (RTO) estabelecido para a workload. Verifique se a restauração de um backup resulta em um recurso contendo os dados originais sem que estejam corrompidos ou inacessíveis e que a perda de dados esteja dentro do objetivo de ponto de recuperação (RPO).

Antipadrões comuns:

- Restaurar um backup, mas não consultar ou recuperar os dados para garantir que a restauração seja útil.
- Presumir a existência de um backup.
- Presumir que o backup de um sistema esteja totalmente operacional e que os dados possam ser recuperados dele.
- Presumir que o tempo para recuperar ou restaurar dados de um backup esteja dentro do RTO para a workload.
- Presumir que os dados contidos no backup estejam dentro do RPO para a workload.
- Restaurar ad hoc, sem usar um runbook, ou o lado de fora de um procedimento automatizado estabelecido.

Benefícios do estabelecimento desta prática recomendada: Testar a recuperação dos backups garante que os dados possam ser restaurados quando necessário, sem a preocupação de que possam estar ausentes ou corrompidos. Os teste também garantem que a restauração e a recuperação sejam possíveis dentro do RTO para a workload e qualquer perda de dados caia dentro do RPO para a workload.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

Testar a capacidade de backup e de restauração aumenta a confiança na aptidão de realizar essas ações durante uma interrupção. Restaure periodicamente os backups em um novo local e execute testes para verificar a integridade dos dados. Alguns testes comuns que devem ser realizados são a verificação

de que todos os dados estão disponíveis, não corrompidos, acessíveis e que qualquer perda de dados está dentro do RPO para a workload. Eles também podem ajudar a verificar se os mecanismos de recuperação são rápidos o suficiente para acomodar o RTO da workload.

1. Identifique as fontes de dados que estão sendo copiados no momento e onde estes backups estão sendo armazenados. Consulte REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes para orientações sobre como implementar isso.
2. Estabeleça critérios para a validação de dados para cada fonte de dados. Diferentes tipos de dados terão propriedades distintas que podem exigir mecanismos de validação diferentes. Considere como validar esses dados antes de se sentir confiante em usá-los na produção.

Algumas maneiras comuns de validar dados são o uso de dados e propriedades de backup, como tipo de dados, formato, soma de verificação, tamanho ou uma combinação deles com lógica de validação personalizada. Por exemplo, pode ser uma comparação dos valores de soma de verificação entre o recurso restaurado e a fonte de dados no momento em que o backup foi criado.

3. Estabeleça um RTO e um RPO para restaurar os dados com base na sua criticidade. Consulte REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados para orientações sobre como implementar isso.
4. Avalie sua capacidade de recuperação. Revise sua estratégia de backup e de restauração para entender se ela pode atender ao RTO e ao RPO e ajuste a estratégia conforme necessário. Com o uso do AWS Resilience Hub, você pode executar uma avaliação da workload. Essa avaliação analisa a aplicação em relação à política de resiliência e relata se as metas de RTO e RPO podem ser atendidas.
5. Faça uma restauração de teste por meio de processos atualmente estabelecidos usados na produção para restauração de dados. Esses processos dependem de como foi feito o backup da fonte de dados original, do formato e do local de armazenamento do próprio backup ou se os dados são reproduzidos de outras fontes. Por exemplo, se você estiver usando um serviço gerenciado, como o AWS Backup, isso poderá ser tão simples quanto restaurar o backup em um novo recurso.. Se usou o AWS Elastic Disaster Recovery, você poderá executar um exercício de recuperação..
6. Valide a recuperação dos dados do recurso restaurado (da etapa anterior) com base nos critérios estabelecidos anteriormente para validação de dados na etapa 2. Os dados restaurados e recuperados contêm o registro ou item mais recente no momento do backup? Esses dados se enquadram no RPO para a workload?
7. Meça o tempo necessário para restauração e recuperação e compare-o ao RTO estabelecido anteriormente na etapa 3. Esse processo se enquadra no RTO para a workload? Por exemplo, compare a data e hora em que o processo de restauração foi iniciado e que a validação da recuperação foi concluída para calcular quanto tempo esse processo leva. Todas as chamadas de API da AWS têm registro de data e hora e essas informações estão disponíveis em AWS CloudTrail. Embora essas informações possam fornecer detalhes sobre o início do processo de restauração, o registro final de data e hora da conclusão da validação deve ser registrado pela lógica de validação. Se estiver usando um processo automatizado, serviços como o Amazon DynamoDB podem ser usado para armazenar essas informações. Além disso, muitos serviços da AWS oferecem um histórico de eventos que fornece informações sobre a data e hora que determinadas ações ocorreram. No AWS Backup, as ações de backup e restauração são

chamadas de Empregos. Eles contêm informações de data e hora como parte dos metadados, que podem ser usados para medir o tempo necessário para restauração e recuperação.

8. Notifique as partes interessadas se a validação de dados falhar ou se o tempo necessário para restauração e recuperação exceder o RTO estabelecido para a workload. Ao implementar a automação para fazer isso, como neste laboratório, é possível usar serviços como o Amazon Simple Notification Service (Amazon SNS) para enviar notificações por push, como e-mail ou SMS, para as partes interessadas. Essas mensagens também podem ser publicadas em aplicativos de mensagens, como o Amazon Chime, o Slack ou o Microsoft Teams, ou usadas para criar tarefas como OpsItems usando o AWS Systems Manager OpsCenter.
9. Automatize este processo para ser executado periodicamente. Por exemplo, serviços como o AWS Lambda ou uma máquina de estado no AWS Step Functions podem ser usados para automatizar os processos de restauração e recuperação, e é possível usar o Amazon EventBridge para acionar esse fluxo de trabalho de automação periodicamente, conforme mostrado no diagrama de arquitetura abaixo. Saiba como Automatizar validação de recuperação de dados com o AWS Backup. Além disso, o laboratório do Well-Architected fornece uma experiência prática em como automatizar várias das etapas aqui.

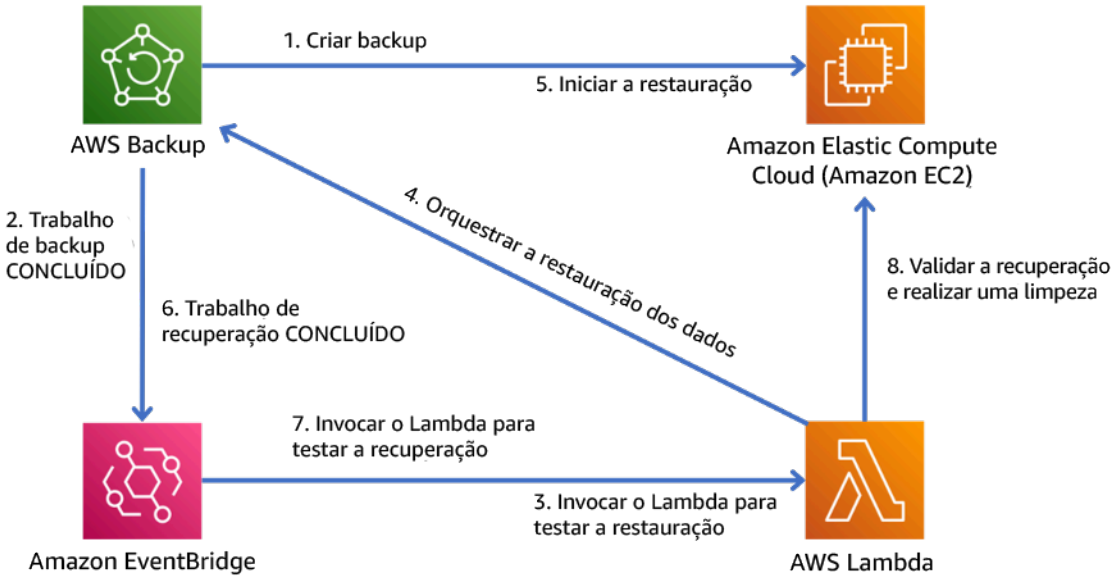


Figura 9. Um processo de backup e restauração automatizado

Nível de esforço para o plano de implementação: De moderado a alto dependendo da complexidade do critério de validação.

Recursos

Documentos relacionados:

- Automatizar validação de recuperação de dados com o AWS Backup
- Parceiro do APN: parceiros que podem ajudar com o backup
- AWS Marketplace: produtos que podem ser usados para backup
- Criar uma regra do EventBridge que é acionada de acordo com uma programação
- Backup e restauração sob demanda para o DynamoDB
- O que é o AWS Backup?
- O que é o AWS Step Functions?
- O que é o AWS Elastic Disaster Recovery
- AWS Elastic Disaster Recovery

Exemplos relacionados:

- Laboratório do Well-Architected: teste de backup e restauração de dados

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.

Práticas recomendadas

- REL10-BP01 Implantar a workload em vários locais
- REL10-BP02 Escolher os locais apropriados para sua implantação de vários locais
- REL10-BP03 Automatizar a recuperação de componentes restritos a um único local
- REL10-BP04 Usar arquiteturas de anteparo para limitar o escopo de impacto

REL10-BP01 Implantar a workload em vários locais

Distribua os dados e os recursos da workload por várias zonas de disponibilidade ou por Regiões da AWS, quando necessário. A diversidade dos locais pode variar conforme a necessidade.

Um dos princípios fundamentais do design de serviço na AWS é evitar pontos únicos de falha em infraestrutura física subjacente. Isso nos motiva a criar software e sistemas que usam várias zonas de disponibilidade e são resilientes à falha de uma única zona. De modo similar, os sistemas são criados para serem resilientes à falha de um único nó de computação, volume de armazenamento ou instância de um banco de dados. Ao criar um sistema que dependa de componentes redundantes, é importante garantir que os componentes operem de modo independente e, no caso de Regiões da AWS, de modo autônomo. Os benefícios obtidos com cálculos teóricos de disponibilidade com componentes redundantes só serão válidos se isso for verdadeiro.

Zonas de disponibilidade (AZ)

As Regiões da AWS são compostas de várias zonas de disponibilidade projetadas para serem independentes umas das outras. Cada zona de disponibilidade é separada por uma grande distância física de outras zonas para evitar cenários de falha correlacionados devido a riscos ambientais, como incêndios, enchentes e tornados. Cada zona de disponibilidade tem uma infraestrutura física independente: conexões dedicadas à rede elétrica, fontes de alimentação de reserva independentes, serviços mecânicos independentes e conectividade de rede independente dentro e além da zona de disponibilidade. O design limita as falhas em qualquer um desses sistemas apenas à AZ afetada. Apesar de estarem geograficamente separadas, as zonas de disponibilidade estão localizadas na mesma área regional, permitindo uma rede de alto throughput e baixa latência. Toda a Região da AWS (em todas as zonas de disponibilidade, consistindo em vários datacenters fisicamente independentes) pode ser tratada como um único destino de implantação lógica para a workload, incluindo a capacidade de replicar dados de forma síncrona (por exemplo, entre bancos de dados). Assim, você pode usar as zonas de disponibilidade em uma configuração ativa/ativa ou ativa/em espera.

As zonas de disponibilidade são independentes e, portanto, a disponibilidade da workload aumenta quando ela é projetada para usar várias zonas. Alguns serviços da AWS (incluindo o plano de dados da instância do Amazon EC2) são implantados como serviços estritamente zonais, compartilhando o destino com a zona de disponibilidade em que estão. No entanto, as instâncias do Amazon EC2 nas outras AZs não serão afetadas e continuarão funcionando. Da mesma forma, se uma falha em uma zona de disponibilidade fizer com que um banco de dados do Amazon Aurora falhe, uma instância do Aurora de réplica de leitura em uma AZ não afetada poderá ser promovida automaticamente para primária. Entretanto, os serviços da AWS regionais (como o Amazon DynamoDB) usam várias zonas de disponibilidade em uma configuração ativa/ativa para atingir as metas de design de disponibilidade para aquele serviço, sem a necessidade de configurar o posicionamento da AZ.

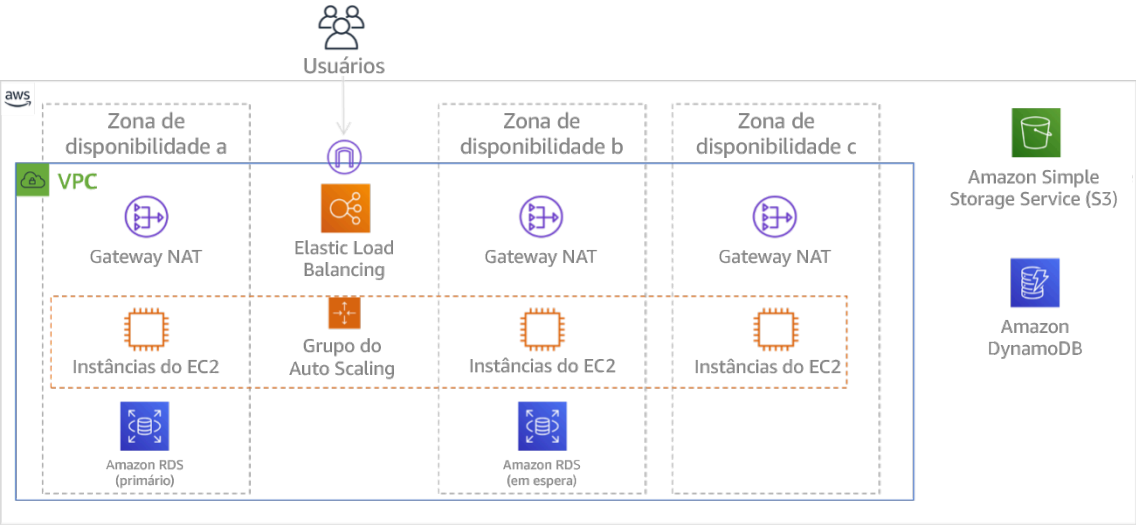


Figura 9: arquitetura multicamadas implantada em três Zonas de disponibilidade. Observe que o Amazon S3 e o Amazon DynamoDB são sempre multi-AZ automaticamente. O ELB também é implantado em todas as três zonas.

Embora os ambientes de gerenciamento da AWS costumem permitir o gerenciamento de recursos dentro de toda a região (várias zonas de disponibilidade), determinados ambientes (incluindo o Amazon EC2 e o Amazon EBS) podem filtrar os resultados para uma única zona de disponibilidade. Quando isso é feito, a solicitação é processada apenas na zona de disponibilidade especificada, o que reduz a exposição a interrupções em outras zonas de disponibilidade. Veja um exemplo da AWS CLI que ilustra como obter informações da instância do Amazon EC2 apenas da zona de disponibilidade us-east-2c:

```
AWS ec2 describe-instances --filters Name=availability-zone,Values=us-east-2c
```

Zonas locais da AWS

Zonas locais da AWS atuam de forma semelhante às zonas de disponibilidade nas suas respectivas Região da AWS, pois elas podem ser selecionadas como um local de posicionamento para recursos zonais da AWS, como sub-redes e instâncias do EC2. O que as torna especiais é que elas estão localizadas não na Região da AWS associada, mas perto de grandes centros populacionais, industriais e de TI onde não existe nenhuma Região da AWS atualmente. No entanto, elas ainda mantêm uma conexão segura e de alta largura de banda entre as workloads locais na zona local e as executadas na Região da AWS. Você deve usar as zonas locais da AWS para implantar workloads mais perto dos seus usuários para requisitos de baixa latência.

Amazon Global Edge Network

A Amazon Global Edge Network consiste em locais da borda em cidades ao redor do mundo. O Amazon CloudFront usa essa rede para entregar conteúdo aos usuários finais com menor latência. O AWS Global Accelerator permite criar endpoints de workload nesses locais da borda para fornecer integração à rede global da AWS próxima aos seus usuários. O Amazon API Gateway habilita endpoints de API otimizados para borda usando uma distribuição do CloudFront para facilitar o acesso do cliente por meio do local da borda mais próximo.

Regiões da AWS

As Regiões da AWS foram projetadas para serem autônomas. Portanto, para usar uma abordagem multirregional, você pode implantar cópias dedicadas de serviços em cada região.

Uma abordagem multirregional é comum para estratégias de recuperação de desastres atenderem aos objetivos de recuperação quando ocorrem eventos pontuais de grande escala. Perceber Planejar para a recuperação de desastres (DR) para obter mais informações sobre essas estratégias. No entanto, aqui focaremos na disponibilidade, que busca entregar um objetivo de tempo de atividade médio ao longo do tempo. Para objetivos de alta disponibilidade, geralmente uma arquitetura multirregional será projetada para ser ativa/ativa, onde cada cópia de serviço (nas suas respectivas regiões) está ativa (atendimento a solicitações).

> **Recomendação**
>
> Os objetivos de disponibilidade para a maioria das workloads podem ser cumpridos usando uma estratégia multi-AZ em uma única Região da AWS. Considere arquiteturas multirregionais somente quando as workloads tiverem requisitos de disponibilidade extrema ou outros objetivos de negócios que exijam uma arquitetura multirregional.

A AWS oferece a capacidade de operar serviços entre regiões. Por exemplo, a AWS fornece replicação contínua e assíncrona de dados usando replicação do Amazon Simple Storage Service (Amazon S3), réplicas de leitura do Amazon RDS (incluindo réplicas de leitura do Aurora) e tabelas globais do Amazon DynamoDB. Com a replicação contínua, as versões dos dados estão disponíveis para uso quase imediato em cada uma das suas regiões ativas.

Ao usar o AWS CloudFormation, você pode definir a infraestrutura e implantá-la de forma consistente em todas as Contas da AWS e Regiões da AWS. O AWS CloudFormation StackSets estende essa funcionalidade, permitindo que criar, atualizar ou excluir pilhas do AWS CloudFormation em várias

contas e regiões com uma única operação. Para implantações de instância do Amazon EC2, uma imagem de máquina da Amazon (AMI) é usada para fornecer informações como configuração de hardware e software instalado. É possível implementar um pipeline do construtor de imagem do Amazon EC2 que cria as AMIs necessárias e as copia para as regiões ativas. Isso garante que essas AMIs de referência (golden) tenham o necessário para implantar e expandir a workload em cada nova região.

Para rotear o tráfego, o Amazon Route 53 e o AWS Global Accelerator permitem a definição de políticas que determinam os usuários que vão para cada endpoint regional ativo. Com o Global Accelerator, você define uma discagem de tráfego para controlar a porcentagem do tráfego que é direcionado para cada endpoint da aplicação. O Route 53 é compatível com a abordagem de porcentagem e com várias outras políticas disponíveis, incluindo as baseadas em geoproximidade e latência. O Global Accelerator aproveita automaticamente a extensa rede de servidores de borda da AWS para integrar o tráfego à estrutura da rede da AWS o mais rápido possível, resultando em menores latências de solicitação.

Todos esses recursos operam de forma a preservar a autonomia de cada região. Há poucas exceções a essa abordagem, incluindo nossos serviços que fornecem entrega global de borda (como o Amazon CloudFront e o Amazon Route 53), juntamente com o ambiente de gerenciamento para o serviço AWS Identity and Access Management (IAM). A maioria dos serviços opera totalmente dentro de uma única região.

Datacenter no local

Para workloads executadas em um datacenter on-premises, arquitete uma experiência híbrida quando possível. O AWS Direct Connect fornece uma conexão de rede dedicada entre o local e a AWS, permitindo que você execute em ambos.

Outra opção é executar a infraestrutura e os serviços da AWS on-premises usando o AWS Outposts. O AWS Outposts é um serviço totalmente gerenciado que estende a infraestrutura da AWS, os serviços da AWS, as APIs e as ferramentas para o seu datacenter. A mesma infraestrutura de hardware usada na Nuvem AWS é instalada no seu datacenter. O AWS Outposts é então conectados à Região da AWS mais próxima. Em seguida, você pode usar AWS Outposts para oferecer suporte a cargas de trabalho com baixa latência ou requisitos de processamento de dados locais.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Use várias zonas de disponibilidade e Regiões da AWS. Distribua os dados e os recursos da workload por várias zonas de disponibilidade ou por Regiões da AWS, quando necessário. A diversidade dos locais pode variar conforme a necessidade.
  - Os serviços regionais são inerentemente implantados nas zonas de disponibilidade.
    - Isso inclui o Amazon S3, o Amazon DynamoDB e o AWS Lambda (quando não conectados a uma VPC).
  - Implemente suas cargas de trabalho baseadas em contêiner, instância e função em várias zonas de disponibilidade. Use datastores multizona, incluindo caches. Use os recursos do EC2 Auto Scaling, o posicionamento de tarefas do ECS, a configuração da função do AWS Lambda ao executá-lo na sua VPC e clusters do ElastiCache.
    - Use sub-redes que estão em zonas de disponibilidade separadas ao implantar grupos de Auto Scaling.
      - Exemplo: distribuição de instâncias entre zonas de disponibilidade
      - Estratégias de posicionamento de tarefas do Amazon ECS
      - Configuração de uma função do AWS Lambda para acessar recursos em uma Amazon VPC
      - Escolha de regiões e zonas de disponibilidade
    - Use sub-redes que estão em zonas de disponibilidade separadas ao implantar grupos de Auto Scaling.
      - Exemplo: distribuição de instâncias entre zonas de disponibilidade
    - Use os parâmetros de posicionamento de tarefas do ECS, especificando grupos de sub-rede do banco de dados.
      - Estratégias de posicionamento de tarefas do Amazon ECS
    - Use sub-redes em várias zonas de disponibilidade ao configurar uma função para executar na sua VPC.
      - Configuração de uma função do AWS Lambda para acessar recursos em uma Amazon VPC
    - Use várias zonas de disponibilidade com os clusters do ElastiCache.
      - Escolha de regiões e zonas de disponibilidade
- Se a workload precisar ser implantada em várias regiões, escolha uma estratégia multirregional. A maioria das necessidades de confiabilidade pode ser atendida em uma única Região da AWS usando uma estratégia de várias zonas de disponibilidade. Use uma estratégia multirregional quando necessário para atender às suas demandas empresariais.

  - AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
    - O backup para outra Região da AWS pode servir como mais uma camada visando garantir que os dados estejam disponíveis quando necessário.
    - Algumas workloads têm requisitos regulamentares que exigem o uso de uma estratégia multirregional.
- Avalie o AWS Outposts para a workload. Se a carga de trabalho exigir baixa latência do datacenter no local ou tiver requisitos de processamento de dados locais. Em seguida, execute a infraestrutura e os serviços da AWS on-premises usando o AWS Outposts
  - O que é o AWS Outposts?
- Determine se as zonas locais da AWS ajudam você a fornecer serviços aos usuários. Se você tiver requisitos de baixa latência, veja se as zonas locais da AWS estão próximas dos seus usuários. Se estiverem, use-as para implantar as workloads mais próximas desses usuários.
  - Perguntas frequentes sobre zonas locais da AWS

Recursos

Documentos relacionados:

- Infraestrutura global da AWS
- Perguntas frequentes sobre zonas locais da AWS
- Estratégias de posicionamento de tarefas do Amazon ECS
- Escolha de regiões e zonas de disponibilidade
- Exemplo: distribuição de instâncias entre zonas de disponibilidade
- Tabelas globais: replicação em várias regiões com o DynamoDB
- Uso de bancos de dados globais do Amazon Aurora
- Série de blogs sobre a criação de uma aplicação multirregional com os serviços da AWS
- O que é o AWS Outposts?

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
- AWS re:Invent 2019: Innovation and operation of the AWS global network infrastructure (NET339)

REL10-BP02 Escolher os locais apropriados para sua implantação de vários locais

Resultado desejado

Para alta disponibilidade, sempre (que possível) implante os componentes da workload em várias zonas de disponibilidade (AZs), conforme mostrado na figura 10. Para workloads com requisitos de resiliência extrema, avalie cuidadosamente as opções para uma arquitetura multirregional.

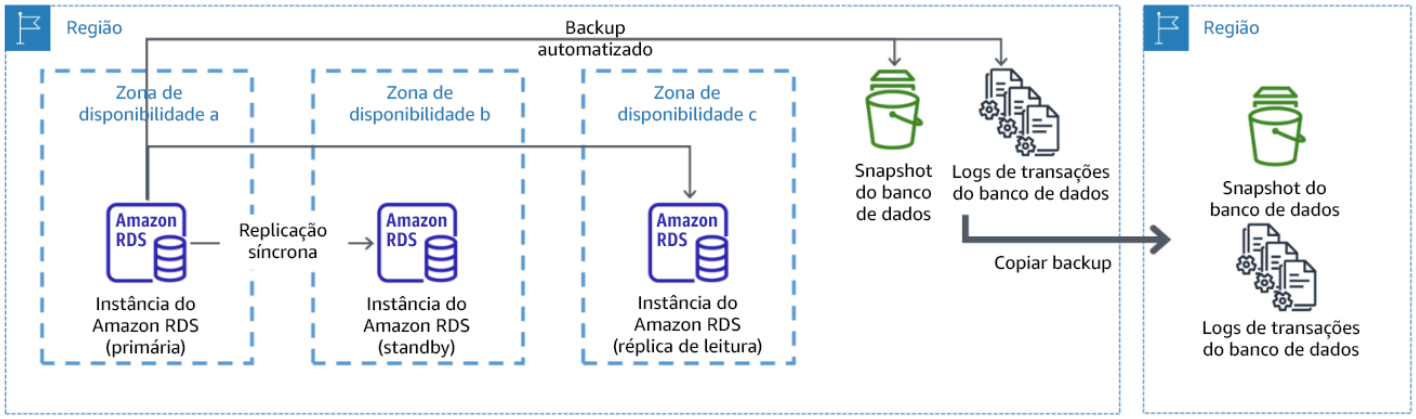


Figura 10: Uma implantação de banco de dados multi-AZ resiliente com backup para outra região da AWS

Antipadrões comuns:

- Projetar uma arquitetura multirregional quando uma arquitetura multi-AZ seria suficiente para atender aos requisitos.
- Não contabilizar as dependências entre os componentes da aplicação caso os requisitos de resiliência e de vários locais forem diferentes entre esses componentes.

Benefícios do estabelecimento desta prática recomendada:

Para resiliência, você deve usar uma abordagem que construa camadas de defesa. Uma camada protege contra interrupções menores e mais comuns criando uma arquitetura altamente disponível usando várias AZs. Outra camada de defesa destina-se a proteger contra eventos raros, como desastres naturais generalizados e interrupções em nível regional. Essa segunda camada envolve arquitetar a aplicação para abranger várias Regiões da AWS.

- A diferença entre as disponibilidades de 99,5% e 99,99% é superior a 3,5 horas por mês. A disponibilidade esperada de uma workload só pode atingir “quatro noves” se estiver em várias AZs.

- Ao executar a workload em várias AZs, é possível isolar falhas de energia, refrigeração, rede e a maioria dos desastres naturais, como incêndio e inundação.
- A implementação de uma estratégia multirregional para a workload ajuda a protegê-la contra desastres naturais generalizados, que afetam uma grande área geográfica de um país, ou falhas técnicas de escopo regional. Esteja ciente de que a implementação de uma arquitetura multirregional pode ser complexa e, geralmente, não é necessária para a maioria das workloads.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

Para um evento de desastre baseado em interrupção ou perda parcial de uma zona de disponibilidade, implementar uma workload altamente disponível em várias zonas de disponibilidade em uma única Região da AWS ajuda a atenuar os desastres naturais e técnicos. Cada Região da AWS é composta por várias zonas de disponibilidade, cada uma isolada de falhas nas outras zonas e separadas por uma distância significativa. No entanto, para um evento de desastre que inclua o risco de perder vários componentes da zona de disponibilidade, distantes umas das outras de forma significativa, deve-se implementar opções de recuperação de desastres para atenuar as falhas de escopo regional. Para workloads que exigem resiliência extrema (infraestrutura crítica, aplicações relacionados à integridade, infraestrutura do sistema financeiro etc.), pode ser necessária uma estratégia multirregional.

Etapas da implementação

1. Avalie a workload e determine se as necessidades de resiliência podem ser atendidas por uma abordagem multi-AZ (Região da AWS única) ou se elas requerem uma abordagem multirregional. A implementação de uma arquitetura multirregional para atender a esses requisitos introduzirá complexidade adicional, portanto, considere cuidadosamente seu caso de uso e seus requisitos. Os requisitos de resiliência quase sempre podem ser atendidos usando uma única Região da AWS. Considere os seguintes requisitos possíveis ao determinar a necessidade de usar várias regiões:
    a. Recuperação de desastres (DR): para um evento de desastre baseado em interrupção ou perda parcial de uma zona de disponibilidade, implementar uma workload altamente disponível em várias zonas de disponibilidade em uma única Região da AWS ajuda a atenuar os desastres naturais e técnicos. Para um evento de desastre que inclua o risco de perder vários componentes da zona de disponibilidade, distantes umas das outras de forma significativa,

deve-se implementar recuperação de desastres multirregional para atenuar os desastres naturais ou as falhas técnicas de escopo regional.

b. Alta disponibilidade (AD): é possível usar uma arquitetura multirregional (usando várias AZs em cada região) para alcançar uma disponibilidade superior a quatro noves (> 99,99%).

c. Localização de pilhas: ao implantar uma workload para um público global, é possível implantar pilhas localizadas em diferentes Regiões da AWS para atender o público nessas regiões. A localização pode incluir idioma, moeda e tipos de dados armazenados.

d. Proximidade aos usuários: ao implantar uma workload para um público global, é possível reduzir a latência implantando pilhas em Regiões da AWS perto de onde os usuários finais estão.

e. Residência de dados: algumas workloads estão sujeitas a requisitos de residência de dados, em que os dados de determinados usuários devem permanecer dentro das fronteiras de um país específico. Com base na regulamentação em questão, você pode optar por implantar uma pilha inteira ou apenas os dados na Região da AWS dentro dessas fronteiras.

2. Veja alguns exemplos de funcionalidade multi-AZ fornecida pelos serviços da AWS:

   a. Para proteger workloads usando o EC2 ou o ECS, implante um Elastic Load Balancer na frente dos recursos de computação. Em seguida, o Elastic Load Balancing fornece a solução para detectar instâncias em zonas com problemas de integridade e rotear o tráfego para as íntegras.

      i. Conceitos básicos do Application Load Balancers

      ii. Conceitos básicos do Network Load Balancers

   b. Em caso de instâncias do EC2 executando software comercial pronto para uso que não oferece suporte ao balanceamento de carga, é possível obter uma forma de tolerância a falhas implementando uma metodologia de recuperação de desastre multi-AZ.

      i. the section called “REL13-BP02 Usar estratégias de recuperação definidas para atender aos objetivos de recuperação”

   c. Para tarefas do Amazon ECS, implante seu serviço uniformemente em três AZs para alcançar um equilíbrio entre disponibilidade e custo.

      i. Práticas recomendadas de disponibilidade do Amazon ECS | Contêineres

   d. Para os que não são Aurora Amazon RDS, você pode escolher multi-AZ como uma opção de configuração. Em caso de falha da instância de banco de dados primário, o Amazon RDS promove automaticamente um banco de dados em espera para receber o tráfego em outra zona de disponibilidade. Também é possível criar réplicas de leitura multirregionais para melhorar a resiliência.

      i. Implantações multi-AZ do Amazon RDS

   ii. Criação de uma réplica de leitura em uma Região da AWS diferente

3. Veja alguns exemplos de funcionalidade multirregional fornecida pelos serviços da AWS:

   a. Para workloads do Amazon S3, em que a disponibilidade multi-AZ é fornecida automaticamente pelo serviço, considere os pontos de acesso multirregionais se for necessária uma implantação multirregional.

      i. Pontos de acesso multirregionais no Amazon S3

   b. Para tabelas do DynamoDB, em que a disponibilidade multi-AZ é fornecida automaticamente pelo serviço, é possível converter tabelas existentes em tabelas globais para aproveitar várias regiões.

      i. Conversão de tabelas de região única do Amazon DynamoDB em tabelas globais

   c. Se a workload for liderada pelo Application Load Balancers ou pelo Network Load Balancers, use o AWS Global Accelerator para melhorar a disponibilidade da aplicação direcionando o tráfego para várias regiões que contenham endpoints íntegros.

      i. Endpoints para aceleradores padrão no AWS Global Accelerator – AWS Global Accelerator (amazon.com)

   d. Para aplicações que utilizam o AWS EventBridge, considere os barramentos entre regiões para encaminhar eventos para outras regiões selecionadas.

      i. Envio e recebimento de eventos do Amazon EventBridge entre Regiões da AWS

   e. Para bancos de dados do Amazon Aurora, considere os bancos de dados globais do Aurora, que abrangem várias regiões da AWS. Os clusters existentes também podem ser modificados para adicionar novas regiões.

      i. Conceitos básicos dos bancos de dados globais do Amazon Aurora

   f. Se a workload incluir chaves de criptografia do AWS Key Management Service (AWS KMS), considere se as chaves multirregionais são apropriadas para a aplicação.

      i. Chaves multirregionais no AWS KMS

   g. Para recursos de outros serviços da AWS, consulte Série sobre a criação de uma aplicação multirregional com os serviços da AWS

Nível de esforço para o plano de implementação: Moderado a alto

Recursos

Documentos relacionados:

- Série sobre a criação de uma aplicação multirregional com os serviços da AWS
- Arquitetura de recuperação de desastres (DR) na AWS, parte IV: multissite ativo-ativo
- Infraestrutura global da AWS
- Perguntas frequentes sobre zonas locais da AWS
- Arquitetura de recuperação de desastres (DR) na AWS, parte I: estratégias de recuperação na nuvem
- Recuperação de desastres é diferente na nuvem
- Tabelas globais: replicação em várias regiões com o DynamoDB

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
- Auth0: arquitetura multirregional de alta disponibilidade que escala a até mais de 1,5 bilhão de logins por mês com failover automático

Exemplos relacionados:

- Arquitetura de recuperação de desastres (DR) na AWS, parte I: estratégias de recuperação na nuvem
- DTCC alcança resiliência muito além do que conseguem em ambiente on-premises
- Expedia Group usa uma arquitetura multirregional e de várias zonas de disponibilidade com um serviço de DNS proprietário para adicionar resiliência às aplicações
- Uber: recuperação de desastres para Kafka multirregional
- Netflix: ativo-ativo para resiliência multirregional
- Como criamos residência de dados para o Atlassian Cloud
- Intuit TurboTax executa em duas regiões

REL10-BP03 Automatizar a recuperação de componentes restritos a um único local

Se os componentes da workload só puderem ser executados em uma única zona de disponibilidade ou em um datacenter on-premises, você deverá implementar capacidade suficiente para fazer uma recompilação completa da workload de acordo com os objetivos de recuperação definidos.

Se a melhor prática para implantar a carga de trabalho em vários locais não for possível devido a restrições tecnológicas, você deverá implementar um caminho alternativo para a resiliência. Você deve automatizar a capacidade de recriar a infraestrutura necessária, reimplantar aplicativos e recriar os dados necessários para esses casos.

Por exemplo, o Amazon EMR executa todos os nós de um determinado cluster na mesma zona de disponibilidade, pois a execução de um cluster na mesma zona melhora a performance dos fluxos de trabalho, pois fornece uma taxa de acesso a dados mais alta. Se esse componente for necessário para a resiliência da workload, você deverá ter uma maneira de reimplantar o cluster e seus dados. Além disso, para o Amazon EMR, você deve provisionar redundância de maneiras diferentes de usar o multi-AZ. Você pode provisionar vários nós. Com o uso do Sistema de arquivos do EMR (EMRFS), os dados no EMR podem ser armazenados no Amazon S3, que podem ser replicados em várias zonas de disponibilidade ou Regiões da AWS.

Da mesma forma, o Amazon Redshift, por padrão, provisiona o cluster em uma zona de disponibilidade escolhida aleatoriamente dentro da Região da AWS selecionada. Todos os nós de cluster são provisionados na mesma zona.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Implemente a autorreparação. Quando possível, use a escalabilidade automática para implantar instâncias ou contêineres. Quando não for possível, use a recuperação automática de instâncias do EC2 ou implemente a automação de autorreparação com base nos eventos de ciclo de vida do contêiner do Amazon EC2 ou do ECS.
  - Use os grupos de Auto Scaling para instâncias e cargas de trabalho de contêiner que não têm requisitos de endereço IP de instância única, endereço IP privado, endereço IP elástico e metadados de instância.
    - O que é o EC2 Auto Scaling?
    - Escalabilidade automática do serviço
      - É possível usar os dados do usuário da configuração de execução para implementar uma automação capaz de fazer a autorreparação da maioria das cargas de trabalho.
  - Use a recuperação automática de instâncias do EC2 para cargas de trabalho que exigem um endereço do ID de instância única, endereço IP privado, endereço IP elástico e metadados de instância.
    - Recupere sua instância.

    - A recuperação automática enviará alertas de status de recuperação para um tópico do SNS quando a falha na instância for detectada.
- Use os eventos de ciclo de vida da instância do EC2 ou os eventos do ECS para automatizar a autorreparação quando a escalabilidade automática ou a recuperação do EC2 não puder ser usada.
    - Ganchos do ciclo de vida do EC2 Auto Scaling
    - Eventos do Amazon ECS
        - Use os eventos para chamar a automação que recuperará seu componente de acordo com a lógica do processo necessária.

Recursos

Documentos relacionados:

- Eventos do Amazon ECS
- Ganchos do ciclo de vida do EC2 Auto Scaling
- Recupere sua instância.
- Escalabilidade automática do serviço
- O que é o EC2 Auto Scaling?

REL10-BP04 Usar arquiteturas de anteparo para limitar o escopo de impacto

Assim como os anteparos de um navio, esse padrão garante que uma falha seja contida em um pequeno subconjunto de solicitações ou clientes para que o número de solicitações prejudicadas seja limitado e a maioria possa continuar sem erros. Geralmente, os anteparos de dados são chamados de partições, enquanto os anteparos de serviços são conhecidos como células.

Em uma arquitetura baseada em células, cada célula é uma instância completa e independente do serviço e tem um tamanho máximo fixo. À medida que a carga aumenta, as cargas de trabalho também aumenta por meio da adição de células. Uma chave de partição é usada no tráfego de entrada para determinar qual célula processará a solicitação. Qualquer falha é contida na única célula em que ela ocorre. Assim, o número de solicitações prejudicadas é limitado, e as outras células continuam sem erros. É importante identificar a chave de partição adequada para minimizar as interações entre células e evitar a necessidade de envolver serviços de mapeamento complexos em cada solicitação. Os serviços que exigem mapeamento complexo acabam apenas transferindo

o problema para os serviços de mapeamento, enquanto os serviços que exigem interações entre células criam dependências entre células (e, assim, reduzem as melhorias de disponibilidade esperadas desse processo).

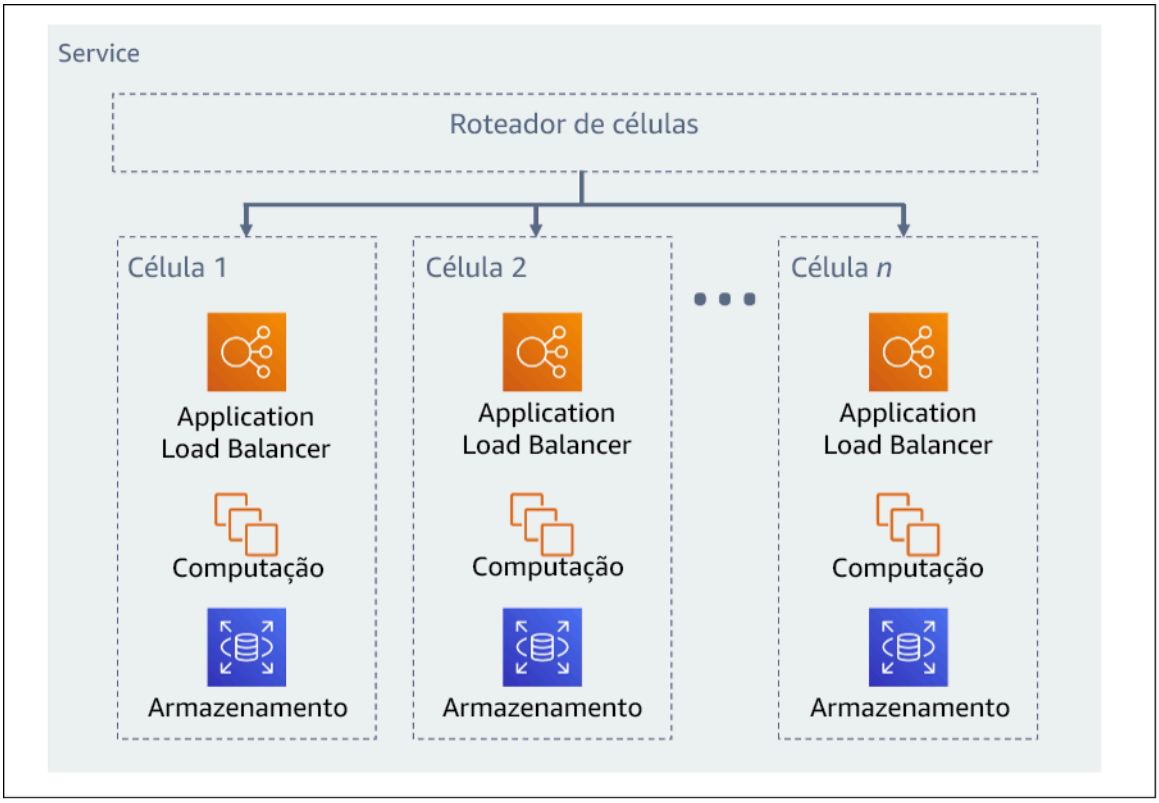


Figura 11: Arquitetura baseada em células

Em sua publicação no blog da AWS, Colm MacCarthaigh explica como o Amazon Route 53 usa o conceito de misturar sharding para isolar as solicitações do cliente em fragmentos. Neste caso, um fragmento consiste em duas ou mais células. Com base na chave de partição, o tráfego de um cliente (ou recursos, ou o que você deseja isolar) é roteado para o fragmento atribuído. No caso de oito células com duas células por fragmento e clientes divididos entre os quatro fragmentos, 25% dos clientes terão impacto no caso de um problema.

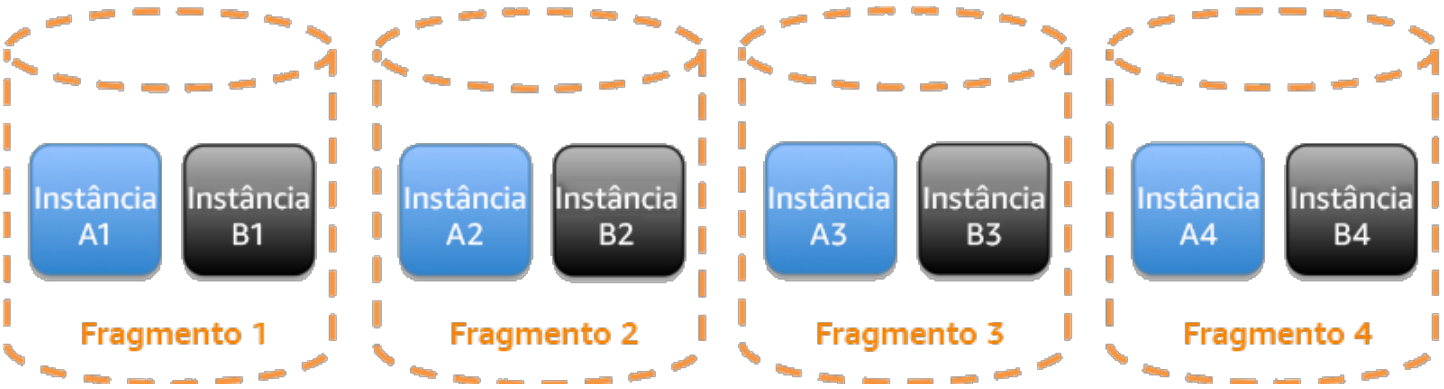


Figura 12: Serviço dividido em quatro fragmentos tradicionais de duas células cada

Com a fragmentação aleatória, você cria fragmentos virtuais de duas células cada e atribui seus clientes a um desses fragmentos virtuais. Quando ocorre um problema, você ainda pode perder um quarto de todo o serviço, mas a maneira como clientes ou recursos são atribuídos significa que o escopo do impacto com fragmentação aleatória é consideravelmente menor que 25%. Com oito células, há 28 combinações exclusivas de duas células, o que significa que há 28 possíveis fragmentos embaralhados (fragmentos virtuais). Se você tiver centenas ou milhares de clientes e atribuir cada cliente a um fragmento embaralhado, o escopo do impacto devido a um problema será apenas 1/28. Isso é sete vezes melhor do que a fragmentação normal.

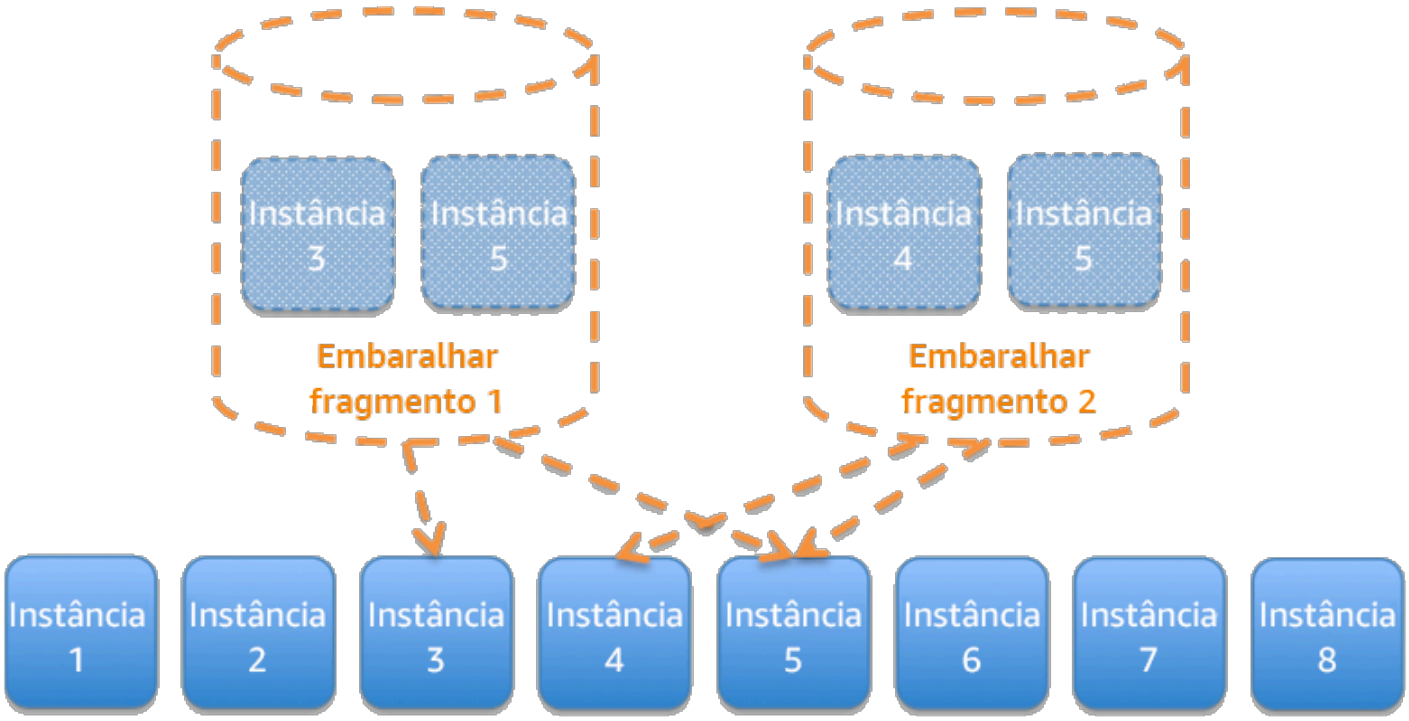


Figura 13: Serviço dividido em 28 fragmentos aleatórios (fragmentos virtuais) de duas células cada (somente dois fragmentos aleatórios dos 28 possíveis são mostrados)

Um fragmento pode ser usado para servidores, filas ou outros recursos, além de células.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- 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, enquanto os anteparos de serviços são conhecidos como células.

  - Laboratório do Well-Architected: isolamento de falhas com fragmentação aleatória
  - Shuffle-sharding: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)
  - AWS re:Invent 2018: How AWS Minimizes the Blast Radius of Failures (ARC338)
- Avalie a arquitetura baseada em células da workload. Em uma arquitetura baseada em células, cada célula é uma instância completa e independente do serviço e tem um tamanho máximo fixo. À medida que a carga aumenta, as cargas de trabalho também aumenta por meio da adição de células. Uma chave de partição é usada no tráfego de entrada para determinar qual célula processará a solicitação. Qualquer falha é contida na única célula em que ela ocorre. Assim, o número de solicitações prejudicadas é limitado, e as outras células continuam sem erros. É importante identificar a chave de partição adequada para minimizar as interações entre células e evitar a necessidade de envolver serviços de mapeamento complexos em cada solicitação. Os serviços que exigem mapeamento complexo acabam apenas transferindo o problema para os serviços de mapeamento, enquanto os serviços que exigem interações entre células reduzem a autonomia das células (e, assim, as melhorias de disponibilidade esperadas desse processo).
  - Em sua publicação no blog da AWS, Colm MacCarthaigh explica como o Amazon Route 53 usa o conceito de fragmentação aleatória para isolar as solicitações do cliente em fragmentos.
    - Fragmentação aleatória: isolamento de falhas massivo e mágico

Recursos

Documentos relacionados:

- Fragmentação aleatória: isolamento de falhas massivo e mágico
- A Amazon Builders' Library: isolamento de workload usando a fragmentação aleatória

Vídeos relacionados:

- AWS re:Invent 2018: How AWS Minimizes the Blast Radius of Failures (ARC338)
- Shuffle-sharding: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)

Exemplos relacionados:

- Laboratório do Well-Architected: isolamento de falhas com fragmentação aleatória

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 Tempo médio até a recuperação (MTTR) devem ser projetadas visando a resiliência.

Práticas recomendadas

- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP02 Failover para recursos íntegros
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL11-BP04 Confiar no plano de dados e não no ambiente de gerenciamento durante a recuperação
- REL11-BP05 Usar a estabilidade estática para evitar o comportamento bimodal
- REL11-BP06 Enviar notificações quando os eventos afetarem a disponibilidade

REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

Monitore constantemente a integridade da workload para que você e seus sistemas automatizados detectem degradações ou falhas assim que elas ocorrerem. Monitore os Key Performance Indicators (KPIs – Indicadores-chave de performance) com base no valor empresarial.

Todos os mecanismos de reparo e recuperação devem começar com a capacidade de detectar problemas rapidamente. As falhas técnicas devem ser detectadas primeiro para que possam ser resolvidas. No entanto, a disponibilidade é baseada na capacidade da workload em entregar valor empresarial, portanto, os indicadores-chave de performance (KPIs) que medem isso precisam fazer parte da sua estratégia de detecção e remediação.

Antipadrões comuns:

- Nenhum alarme foi configurado, portanto as interrupções ocorrem sem notificação.
- Os alarmes existem, mas com limites que não permitem um tempo adequado para reação.
- As métricas não são coletadas com frequência suficiente para atender ao Recovery Time Objective (RTO – Objetivo do tempo de recuperação).
- Dentre os níveis da carga de trabalho, somente aquele voltado ao cliente é monitorado ativamente.
- Coleta apenas das métricas técnicas, não das métricas de função de negócios.
- Não há métricas que medem a experiência do usuário da carga de trabalho.

Benefícios do estabelecimento dessa prática recomendada: O monitoramento adequado de todas as camadas reduz o tempo de detecção e, assim, permite reduzir o tempo de recuperação.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Determine o intervalo de coleta dos componentes com base nas suas metas de recuperação.
  - O intervalo de monitoramento depende da rapidez com que você precisa fazer a recuperação. O tempo de recuperação é determinado pelo tempo necessário para a recuperação. Desse modo, você deve considerar esse tempo e o RTO para determinar a frequência da coleta.
- Configure o monitoramento detalhado dos componentes.
  - Determine se o monitoramento detalhado das instâncias do EC2 e do Auto Scaling é necessário. O monitoramento detalhado fornece métricas de intervalo de 1 minuto, e o monitoramento padrão fornece métricas de intervalo de 5 minutos.
    - Habilitar ou desabilitar o monitoramento detalhado de instância
    - Monitoramento de grupos do Auto Scaling e instâncias usando o Amazon CloudWatch
  - Determine se o monitoramento avançado para RDS é necessário. O monitoramento avançado usa um agente nas instâncias do RDS para obter informações úteis sobre processos ou threads diferentes em uma instância do RDS.
    - Enhanced Monitoring
- Crie métricas personalizadas para medir os indicadores-chave de performance (KPIs) de negócios. As cargas de trabalho implementam as principais funções de negócios. Essas funções devem ser usadas como KPIs que ajudam a identificar quando ocorre um problema indireto.
  - Publicar métricas personalizadas
- Use os canários de usuário para monitorar falhas na experiência do usuário. O teste de transações sintéticas (também conhecido como teste canário, que não deve ser confundido com as implantações canário) que pode executar e simular o comportamento do cliente está entre os processos de teste mais importantes. Execute esses testes constantemente nos endpoints da carga de trabalho de diversos locais remotos.
  - O Amazon CloudWatch Synthetics permite criar canários de usuário
- Crie métricas personalizadas que acompanham a experiência do usuário. Se você puder estabelecer instrumentos de medição da experiência do cliente, conseguirá determinar o momento de degradação da experiência do consumidor.
  - Publicar métricas personalizadas

- Defina alarmes para detectar quando uma parte da carga de trabalho não estiver funcionando corretamente e indicar quando deve ser feita a escalabilidade automática dos recursos. É possível exibir os alarmes em painéis, enviar alertas pelo Amazon SNS ou por e-mail e trabalhar com o Auto Scaling para aumentar ou reduzir a escala verticalmente dos recursos de uma workload.
  - Uso de alarmes do Amazon CloudWatch
- Crie painéis para visualizar as métricas. É possível usar os painéis para ver as tendências, os casos atípicos e outros indicadores de possíveis problemas ou para obter uma indicação de problemas a serem investigados.
  - Uso de painéis do CloudWatch

Recursos

Documentos relacionados:

- O Amazon CloudWatch Synthetics permite criar canários de usuário
- Habilitar ou desabilitar o monitoramento detalhado de instância
- Enhanced Monitoring
- Monitoramento de grupos do Auto Scaling e instâncias usando o Amazon CloudWatch
- Publicar métricas personalizadas
- Uso de alarmes do Amazon CloudWatch
- Uso de painéis do CloudWatch

Exemplos relacionados:

- Laboratório do Well-Architected: nível 300: implementação de verificações de integridade e do gerenciamento de dependências para melhorar a confiabilidade

REL11-BP02 Failover para recursos íntegros

Verifique se, caso ocorra uma falha de recurso, os recursos íntegros podem continuar atendendo às solicitações. Para falhas de localização (como zona de disponibilidade ou Região da AWS), garanta que você tenha sistemas implementados para executar failover para recursos íntegros em locais sem problemas.

Os serviços da AWS, como o Elastic Load Balancing e o AWS Auto Scaling, ajudam a distribuir carga entre recursos e zonas de disponibilidade. Portanto, a falha de um recurso individual (como uma

instância do EC2) ou o comprometimento de uma zona de disponibilidade podem ser atenuados desviando o tráfego para os recursos íntegros restantes. Para as cargas de trabalho multirregionais, o procedimento é mais complicado. Por exemplo, 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 primária e apontar o tráfego para ela em caso de failover. O Amazon Route 53 e o AWS Global Accelerator ajudam a rotear o tráfego entre Regiões da AWS.

Se a workload estiver usando serviços da AWS, como o Amazon S3 ou o Amazon DynamoDB, ela será implantada automaticamente em várias zonas de disponibilidade. Em caso de falha, o ambiente de gerenciamento da AWS roteia automaticamente o tráfego para locais íntegros. Os dados são armazenados de forma redundante em várias zonas de disponibilidade e permanecem disponíveis. Para o Amazon RDS, você deve escolher multi-AZ como opção de configuração e, em caso de falha, a AWS direcionará automaticamente o tráfego para a instância íntegra. Para instâncias do Amazon EC2, tarefas do Amazon ECS ou pods do Amazon EKS, você escolhe em quais zonas de disponibilidade implantar. Em seguida, o Elastic Load Balancing fornece a solução para detectar instâncias em zonas com problemas de integridade e rotear o tráfego para as zonas íntegras. O Elastic Load Balancing pode até mesmo rotear o tráfego para componentes no seu datacenter on-premises.

Para abordagens multirregionais (que também podem incluir datacenters on-premises), o Amazon Route 53 oferece uma maneira de definir domínios da Internet e de atribuir políticas de roteamento que podem incluir verificações de integridade para garantir que o tráfego seja roteado para regiões íntegras. Como alternativa, o AWS Global Accelerator fornece endereços IP estáticos que atuam como um ponto de entrada fixo para a aplicação e, em seguida, roteia para endpoints nas Regiões da AWS de sua escolha, usando a rede global da AWS em vez da Internet para obter melhor performance e confiabilidade.

A AWS aborda o design dos nossos serviços pensando na recuperação de falha. Projetamos os serviços para minimizar o tempo para recuperação de falhas e o impacto sobre os dados. Nossos serviços usam principalmente repositórios de dados que reconhecem solicitações apenas após serem armazenadas de modo durável entre várias réplicas em uma região. Esses serviços e recursos incluem o Amazon Aurora, instâncias de banco de dados multi-AZ doAmazon Relational Database Service (Amazon RDS), o Amazon S3, o Amazon DynamoDB, o Amazon Simple Queue Service (Amazon SQS) e o Amazon Elastic File System (Amazon EFS). Eles são criados para usar isolamento com base em células e usar o isolamento de falhas fornecido por Zonas de disponibilidade. Usamos automação amplamente em nossos procedimentos operacionais. Também otimizamos nossa funcionalidade de substituir e reiniciar para a recuperação rápida de interrupções.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Faça failover para recursos íntegros. Verifique se, caso ocorra uma falha de recurso, os recursos íntegros podem continuar atendendo às solicitações. Para falhas de localização (como zona de disponibilidade ou Região da AWS), garanta que você tenha sistemas implementados para executar failover para recursos íntegros em locais sem problemas.
  - Se a workload estiver usando serviços da AWS, como o Amazon S3 ou o Amazon DynamoDB, ela será implantada automaticamente em várias zonas de disponibilidade. Em caso de falha, o ambiente de gerenciamento da AWS roteia automaticamente o tráfego para locais íntegros.
  - Para o Amazon RDS, você deve escolher multi-AZ como opção de configuração e, em caso de falha, a AWS direcionará automaticamente o tráfego para a instância íntegra.
    - Alta disponibilidade (multi-AZ) para o Amazon RDS
  - Para instâncias do Amazon EC2 ou tarefas do Amazon ECS, você escolhe em quais Zonas de disponibilidade implantar. Em seguida, o Elastic Load Balancing fornece a solução para detectar instâncias em zonas com problemas de integridade e rotear o tráfego para as zonas íntegras. O Elastic Load Balancing pode até mesmo rotear o tráfego para componentes no seu datacenter on-premises.
  - Para abordagens em várias regiões (que também podem incluir datacenters on-premises), certifique-se de que os dados e os recursos de locais íntegros possam continuar atendendo a solicitações
    - Por exemplo, 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 primária e apontar o tráfego para ela em caso de falha no local primário.
      - Visão geral das réplicas de leitura do Amazon RDS
    - O Amazon Route 53 oferece uma maneira de definir domínios da Internet e de atribuir políticas de roteamento que podem incluir verificações de integridade para garantir que o tráfego seja roteado para regiões íntegras. Como alternativa, o AWS Global Accelerator fornece endereços IP estáticos que atuam como um ponto de entrada fixo para a aplicação e, em seguida, roteia para endpoints nas Regiões da AWS de sua escolha, usando a rede global da AWS em vez da Internet pública para obter melhor performance e confiabilidade.
      - Amazon Route 53: escolher uma política de roteamento
      - O que é o AWS Global Accelerator?

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar na automação da sua tolerância a falhas
- AWS Marketplace: produtos que podem ser usados para tolerância a falhas
- AWS OpsWorks: como usar a autorreparação para substituir instâncias com falha
- Amazon Route 53: escolher uma política de roteamento
- Alta disponibilidade (multi-AZ) para o Amazon RDS
- Visão geral das réplicas de leitura do Amazon RDS
- Estratégias de posicionamento de tarefas do Amazon ECS
- Criação de grupos de Auto Scaling do Kubernetes para várias zonas de disponibilidade
- O que é o AWS Global Accelerator?

Exemplos relacionados:

- Laboratório do Well-Architected: nível 300: implementação de verificações de integridade e do gerenciamento de dependências para melhorar a confiabilidade

REL11-BP03 Automatizar a reparação em todas as camadas

Após a detecção de uma falha, use recursos automatizados para executar ações de correção.

Capacidade de reiniciar é uma ferramenta importante para corrigir falhas. Como discutido anteriormente para sistemas distribuídos, uma melhor prática é tornar os serviços sem estado sempre que possível. Isso evita a perda de dados ou a disponibilidade na reinicialização. Na nuvem, você pode (e geralmente deve) substituir todo o recurso (por exemplo, instância do EC2 ou função do Lambda) como parte da reinicialização. A reinicialização em si é uma maneira simples e confiável de se recuperar de falhas. Muitos tipos diferentes de falhas ocorrem em cargas de trabalho. As falhas podem ocorrer em hardware, software, comunicações e operações. Em vez de criar mecanismos novos para capturar, identificar e corrigir cada um dos diferentes tipos de falhas, mapeie várias categorias diferentes de falhas para a mesma estratégia de recuperação. Uma instância pode falhar devido a uma falha de hardware, um bug no sistema operacional, vazamento de memória ou outras causas. Em vez de criar uma correção personalizada para cada situação, trate qualquer uma delas como uma falha de instância. Encerre a instância e permita que o AWS Auto Scaling a substitua. Posteriormente, você pode executar a análise do recurso com falha fora de banda.

Outro exemplo é a capacidade de reiniciar uma solicitação de rede. Aplique a mesma abordagem de recuperação tanto a um tempo limite de rede quanto a uma falha de dependência em que a dependência retorna um erro. Ambos os eventos têm um efeito similar sobre o sistema, assim, em vez de tentar tornar qualquer um dos eventos um “caso especial”, aplique uma estratégia similar de nova tentativa limitada com recuo e variação exponenciais.

Capacidade de reiniciar é um mecanismo de recuperação destacado em computação orientada à recuperação (ROC) e arquiteturas de cluster de alta disponibilidade.

É possível usar o Amazon EventBridge para monitorar e filtrar eventos, como alarmes do CloudWatch ou alterações no estado de outros serviços da AWS. Com base nas informações do evento, ele pode acionar o AWS Lambda, o AWS Systems Manager Automation ou outros destinos para executar a lógica de correção personalizada na workload.

O Amazon EC2 Auto Scaling pode ser configurado para verificar a integridade da instâncias do EC2. Se a instância estiver em qualquer estado que não seja em execução, ou se o status do sistema for prejudicado, o Amazon EC2 Auto Scaling considerará que essa instância não é íntegra e executará uma instância de substituição. Se estiver usando o AWS OpsWorks, você poderá configurar a autorreparação das instâncias do EC2 no nível da camada do OpsWorks.

Para substituições em grande escala (como a perda de uma Zona de disponibilidade inteira), a estabilidade estática é preferida para alta disponibilidade, em vez de tentar obter vários novos recursos de uma só vez.

Antipadrões comuns:

- Implantação de aplicações em instâncias ou contêineres individualmente.
- Implantação de aplicações que não podem ser implantadas em vários locais sem usar a recuperação automática.
- Reparação manual de aplicações que não são reparadas por meio da escalabilidade e recuperação automáticas.

Benefícios do estabelecimento desta prática recomendada: A reparação automatizada, mesmo que a carga de trabalho só possa ser implantada em um local por vez, reduzirá o tempo médio até a recuperação e garantirá a disponibilidade da carga de trabalho.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Use os grupos de Auto Scaling para implantar níveis em uma workload. O Auto Scaling pode executar a autorreparação em aplicativos sem estado e adicionar e remover capacidade.
  - Como funciona o AWS Auto Scaling
- Implemente a recuperação automática em instâncias do EC2 que tenham aplicativos implantados que não possam ser implantados em vários locais e possam tolerar a reinicialização em caso de falhas. É possível usar a recuperação automática para substituir o hardware com falha e reiniciar a instância quando o aplicativo não puder ser implantado em vários locais. Os metadados e os endereços IP associados da instância são mantidos, assim como os volumes e pontos de montagem do Amazon EBS para o Elastic File Systems ou File Systems for Lustre e Windows.
  - Recuperação automática do Amazon EC2
  - Amazon Elastic Block Store (Amazon EBS)
  - Amazon Elastic File System (Amazon EFS)
  - O que é o Amazon FSx for Lustre?
  - O que é o Amazon FSx for Windows File Server?
    - Ao usar o AWS OpsWorks, é possível configurar a autorreparação das instâncias do EC2 no nível da camada
      - AWS OpsWorks: como usar a autorreparação para substituir instâncias com falha
- Implemente a recuperação automatizada usando o AWS Step Functions e o AWS Lambda quando não for possível usar a escalabilidade ou a recuperação automáticas, ou quando a recuperação automática falhar. Quando não for possível usar a escalabilidade ou a recuperação automáticas ou quando a recuperação automática falhar, você poderá automatizar a reparação usando o AWS Step Functions e o AWS Lambda.
  - O que é o AWS Step Functions?
  - O que é o AWS Lambda?
    - É possível usar o Amazon EventBridge para monitorar e filtrar eventos, como alarmes do CloudWatch ou alterações no estado de outros serviços da AWS. Com base nas informações do evento, ele pode acionar o AWS Lambda (ou outros destinos) para executar a lógica de correção personalizada na workload.
      - O que é o Amazon EventBridge?
      - Uso de alarmes do Amazon CloudWatch

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar na automação da sua tolerância a falhas
- AWS Marketplace: produtos que podem ser usados para tolerância a falhas
- AWS OpsWorks: como usar a autorreparação para substituir instâncias com falha
- Recuperação automática do Amazon EC2
- Amazon Elastic Block Store (Amazon EBS)
- Amazon Elastic File System (Amazon EFS)
- Como funciona o AWS Auto Scaling
- Uso de alarmes do Amazon CloudWatch
- O que é o Amazon EventBridge?
- O que é o AWS Lambda?
- AWS Systems Manager Automation
- O que é o AWS Step Functions?
- O que é o Amazon FSx for Lustre?
- O que é o Amazon FSx for Windows File Server?

Vídeos relacionados:

- Static stability in AWS: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)

Exemplos relacionados:

- Laboratório do Well-Architected: nível 300: implementação de verificações de integridade e do gerenciamento de dependências para melhorar a confiabilidade

REL11-BP04 Confiar no plano de dados e não no ambiente de gerenciamento durante a recuperação

O ambiente de gerenciamento é usado para configurar recursos, e o plano de dados fornece serviços. Os planos de dados geralmente têm metas de design de disponibilidade mais altas do que os ambientes de gerenciamento e costumam ser menos complexos. Ao implementar respostas

de recuperação ou mitigação para eventos que possam ter um impacto na resiliência, o uso de operações do ambiente de gerenciamento pode diminuir a resiliência geral da sua arquitetura. Por exemplo, você pode confiar no plano de dados do Amazon Route 53 para rotear consultas ao DNS com base nas verificações de integridade. Porém, a atualização das políticas de roteamento do Route 53 usa o ambiente de gerenciamento, portanto, não conte com ele para a recuperação.

Os planos de dados do Route 53 respondem as consultas ao DNS, além de realizarem e avaliarem verificações de integridade. Eles são distribuídos globalmente e projetados para um Acordo de Serviço (SLA) de 100% de disponibilidade. As APIs e consoles de gerenciamento do Route 53 usados para criar, atualizar e excluir recursos do Route 53 são executados em ambientes de gerenciamento projetados para priorizar a consistência e a durabilidade necessária para gerenciar o DNS. Para que isso aconteça, os ambientes de gerenciamento estão localizados em uma única região, US East (N. Virginia). Embora ambos os sistemas sejam construídos para serem muito confiáveis, os ambientes de gerenciamento não estão incluídos no SLA. Pode ser que ocorram raros eventos onde o design resiliente do plano de dados permita que ele mantenha a disponibilidade, enquanto os ambientes de gerenciamento não. Para mecanismos de recuperação de desastres e failover, use funções de plano de dados para fornecer a melhor confiabilidade possível.

Para obter mais informações sobre planos de dados, ambientes de gerenciamento e como a AWS cria serviços para atender destinos de alta disponibilidade, consulte o documento estabilidade estática usando Zonas de disponibilidade e a Amazon Builders' Library.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Confie no plano de dados e não no ambiente de gerenciamento ao usar o Amazon Route 53 para recuperação de desastres. O Route 53 Application Recovery Controller ajuda a gerenciar e coordenar o failover usando verificações de prontidão e controles de roteamento. Esses recursos monitoram continuamente a capacidade da aplicação de se recuperar de falhas, permitindo que você controle a recuperação da aplicação em várias Regiões da AWS, zonas de disponibilidade e ambientes on-premises.
  - O que é o Route 53 Application Recovery Controller
  - Criação de mecanismos de recuperação de desastres usando o Amazon Route 53
  - Criação de aplicações altamente resilientes usando o Amazon Route 53 Application Recovery Controller, parte 1: pilha de região única
  - Criação de aplicações altamente resilientes usando o Amazon Route 53 Application Recovery Controller, parte 2: pilha multirregional

- Compreenda quais operações estão no plano de dados e quais estão no ambiente de gerenciamento.
  - Amazon Builders' Library: evite a sobrecarga em sistemas distribuídos colocando o menor serviço no controle
  - API do Amazon DynamoDB (ambiente de gerenciamento e plano de dados)
  - Execuções do AWS Lambda (divididas entre o ambiente de gerenciamento e o plano de dados)
  - Execuções do AWS Lambda (divididas entre o ambiente de gerenciamento e o plano de dados)

Recursos

Documentos relacionados:

- Parceiro da APN: parceiros que podem ajudar na automação da sua tolerância a falhas
- AWS Marketplace: produtos que podem ser usados para tolerância a falhas
- Amazon Builders' Library: evite a sobrecarga em sistemas distribuídos colocando o menor serviço no controle
- API do Amazon DynamoDB (ambiente de gerenciamento e plano de dados)
- Execuções do AWS Lambda (divididas entre o ambiente de gerenciamento e o plano de dados)
- Plano de dados do AWS Elemental MediaStore
- Criação de aplicações altamente resilientes usando o Amazon Route 53 Application Recovery Controller, parte 1: pilha de região única
- Criação de aplicações altamente resilientes usando o Amazon Route 53 Application Recovery Controller, parte 2: pilha multirregional
- Criação de mecanismos de recuperação de desastres usando o Amazon Route 53
- O que é o Route 53 Application Recovery Controller

Exemplos relacionados:

- Introdução ao Amazon Route 53 Application Recovery Controller

REL11-BP05 Usar 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.

A estabilidade estática para implantação de computação (como instâncias ou contêineres do EC2) resultará na mais alta confiabilidade. Isso deve ser ponderado em relação a preocupações de custo. É mais barato provisionar menos capacidade computacional e contar com a execução de novas instâncias em caso de falha. No entanto, para falhas em grande escala (como uma falha de zona de disponibilidade), essa abordagem é menos eficaz porque depende de reagir a prejuízos à medida que ocorrem, em vez de estar preparada para essas deficiências antes que elas ocorram. Sua solução deve ponderar a confiabilidade em comparação com as necessidades de custo para sua carga de trabalho. Ao usar mais zonas de disponibilidade, a quantidade de computação adicional necessária para a estabilidade estática diminui.

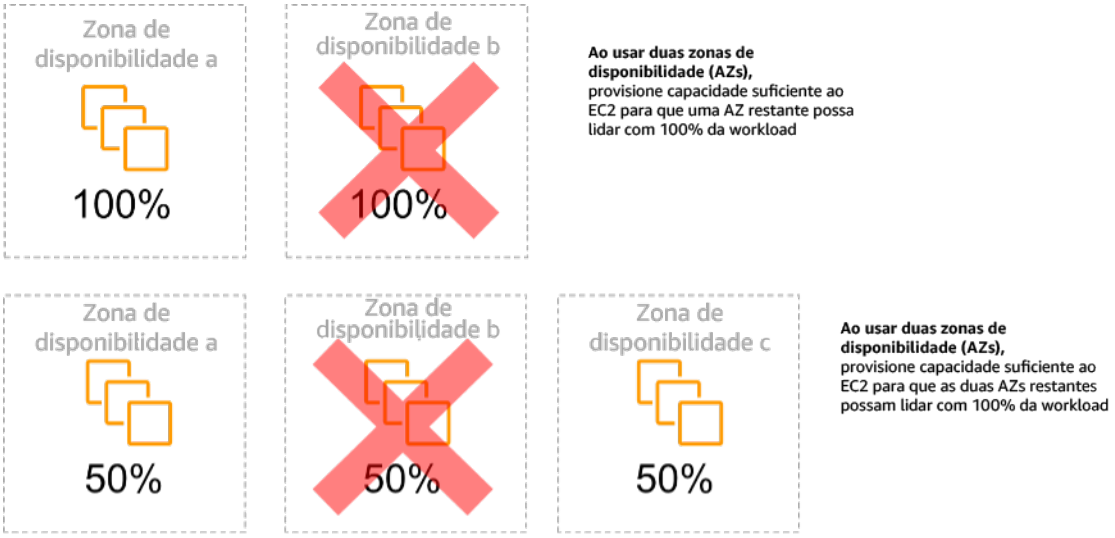


Figura 14: Estabilidade estática de instâncias do EC2 em várias zonas de disponibilidade

Depois que o tráfego for deslocado, use o AWS Auto Scaling para substituir de forma assíncrona instâncias da zona com falha e executá-las nas zonas íntegras.

Outro exemplo de comportamento bimodal seria um tempo limite de rede que poderia fazer com que um sistema tentasse atualizar o estado de configuração de todo o sistema. Isso adicionaria uma carga inesperada a outro componente e pode fazê-lo falhar, levando a outras consequências inesperadas. Esse ciclo de comentário negativo afeta a disponibilidade de sua carga de trabalho. Em vez disso, você deve criar sistemas estaticamente estáveis e operar em apenas um modo. Um design estático estável seria fazer um trabalho constante e sempre atualizar o estado da

configuração em um ritmo fixo. Quando uma chamada falha, a carga de trabalho usa o valor armazenado em cache anteriormente e aciona um alarme.

Outro exemplo de comportamento bimodal é permitir que os clientes ignorem o cache da carga de trabalho em caso de falhas. Isso pode parecer ser uma solução que acomoda as necessidades do cliente, mas não deve ser permitida porque altera significativamente as demandas em sua carga de trabalho e provavelmente resultará em falhas.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- 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.
  - Minimizar dependências em um plano de recuperação de desastres
  - A Amazon Builders' Library: estabilidade estática usando zonas de disponibilidade
  - Static stability in AWS: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)
    - Em vez disso, você deve criar sistemas que sejam estaticamente estáveis e que operem em apenas um modo. Nesse caso, provisione instâncias suficientes em cada zona para processar a workload 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.
    - Outro exemplo de comportamento bimodal é permitir que os clientes ignorem o cache da carga de trabalho em caso de falhas. Isso pode parecer uma solução para acomodar as necessidades do cliente, mas não deve ser permitido porque altera significativamente as demandas em sua carga de trabalho e pode resultar em falhas.

Recursos

Documentos relacionados:

- Minimizar dependências em um plano de recuperação de desastres
- A Amazon Builders' Library: estabilidade estática usando zonas de disponibilidade

Vídeos relacionados:

- Static stability in AWS: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)

REL11-BP06 Enviar 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.

A correção automatizada permite que a carga de trabalho seja confiável. No entanto, ele também pode obscurecer problemas subjacentes que precisam ser resolvidos. Implemente eventos e monitoramento apropriados para que você possa detectar padrões de problemas, incluindo aqueles abordados pela correção automática, para que você possa resolver problemas de causa raiz. Alarmes do Amazon CloudWatch podem ser acionados com base em falhas que ocorrem. Eles também podem ser acionados com base em ações de correção automatizadas executadas. Alarmes do CloudWatch podem ser configurados para enviar e-mails ou registrar incidentes em sistemas de rastreamento de incidentes de terceiros usando a integração com o Amazon SNS.

Antipadrões comuns:

- Envio de alarmes sem necessidade de reação.
- Execução da automação de autorreparação, mas sem notificar que a reparação era necessária.

Benefícios do estabelecimento dessa prática recomendada: As notificações de eventos de recuperação garantem que você não ignore problemas que ocorrem com pouca frequência.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Alarmes de indicadores-chave de performance de negócios quando eles excedem um limite baixo. Possuir um alarme de limite baixo nos KPIs de negócios ajuda a saber quando a workload está indisponível ou não funcional.
  - Criação de um alarme do CloudWatch com base em um limite estático
- Alarmes de eventos que invocam automação de reparação. Você pode invocar diretamente uma API do SNS para enviar notificações com qualquer automação criada.
  - O que é o Amazon Simple Notification Service?

Recursos

Documentos relacionados:

- Criação de um alarme do CloudWatch com base em um limite estático

- O que é o Amazon EventBridge?
- O que é o Amazon Simple Notification Service?

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.

Práticas recomendadas

- REL12-BP01 Usar playbooks para investigar falhas
- REL12-BP02 Realizar análise pós-incidente
- REL12-BP03 Testar os requisitos funcionais
- REL12-BP04 Testar os requisitos de escalabilidade e performance
- REL12-BP05 Testar a resiliência por meio da engenharia do caos
- REL12-BP06 Realizar dias de jogo regularmente

REL12-BP01 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.

O playbook é um planejamento proativo que você deve fazer para poder executar ações reativas com eficácia. Quando cenários de falha não cobertos pelo playbook forem encontrados na produção, resolva primeiro o problema (apague o fogo). Em seguida, volte e veja as etapas que você seguiu para resolver o problema e use-as para adicionar uma nova entrada no playbook.

Observe que playbooks são usados em resposta a incidentes específicos, enquanto runbooks são usados para alcançar resultados específicos. Muitas vezes, runbooks são usados para atividades de rotina e os playbooks são usados para responder a eventos que não são rotineiros.

Antipadrões comuns:

- Planejar a implantação de uma carga de trabalho sem conhecer os processos para diagnosticar problemas ou responder a incidentes.

- Decisões não planejadas de quais sistemas coletar logs e métricas ao investigar um evento.
- Não armazenar as métricas e os eventos pelo tempo suficiente para recuperar os dados.

Benefícios do estabelecimento desta prática recomendada: Capturar playbooks garante que os processos possam ser seguidos de forma consistente. A codificação dos seus playbooks limita a introdução de erros por atividades manuais. A automação dos playbooks reduz o tempo de resposta a um evento ao eliminar a necessidade de intervenção de membros da equipe ou ao fornecer a eles informações adicionais desde o início da intervenção.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Use playbooks para identificar problemas. Os manuais são processos documentados para investigar problemas. Faça a documentação dos processos em playbooks para permitir respostas consistentes e rápidas em cenários de falha. Os playbooks devem incluir as informações e as diretrizes necessárias para que uma pessoa com as devidas qualificações colete as informações aplicáveis, identifique possíveis fontes de falha, isole as falhas e determine os fatores contribuintes (realize uma análise pós-incidente).
  - Implemente playbooks como código. Execute suas operações como código ao criar scripts de seus playbooks para garantir a consistência e reduzir os erros causados por processos manuais. Os playbooks podem ser compostos por vários scripts representando as diferentes etapas que podem ser necessárias para identificar os fatores que contribuem para um problema. As atividades do runbook podem ser acionadas ou executadas como parte das atividades do playbook, ou podem solicitar a execução de um playbook em resposta a eventos identificados.
    - Automatizar playbooks operacionais com o AWS Systems Manager
    - AWS Systems Manager Run Command
    - AWS Systems Manager Automation
    - O que é o AWS Lambda?
    - O que é o Amazon EventBridge?
    - Usar alarmes do Amazon CloudWatch

Recursos

Documentos relacionados:

- AWS Systems Manager Automation
- AWS Systems Manager Run Command
- Automatizar playbooks operacionais com o AWS Systems Manager
- Usar alarmes do Amazon CloudWatch
- Uso de canários (Amazon CloudWatch Synthetics)
- O que é o Amazon EventBridge?
- O que é o AWS Lambda?

Exemplos relacionados:

- Automating operations with Playbooks and Runbooks (Automatização de operações com manuais e runbooks)

REL12-BP02 Realizar 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.

Avalie por que os testes existentes não encontraram o problema. Adicione testes para esse caso se os testes ainda não existirem.

Antipadrões comuns:

- Encontrar fatores contribuintes, mas não continuar buscando mais profundamente outros possíveis problemas e abordagens de mitigação.
- Identificar apenas as causas de erros humanos e não oferecer nenhum treinamento ou automação que possa evitar erros humanos.

Benefícios do estabelecimento dessa prática recomendada: A realização de análises pós-incidentes e o compartilhamento dos resultados permitem que outras cargas de trabalho atenuem o risco caso tenham implementado os mesmos fatores contribuintes, além de permitir que elas implementem a mitigação ou a recuperação automatizada antes que ocorra um incidente.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Estabeleça um padrão para sua análise pós-incidente. Uma boa análise pós-incidente oferece oportunidades para propor soluções comuns a problemas com padrões de arquitetura usados em outros locais nos sistemas.
  - Garantir que os fatores contribuintes sejam justos e isentos de acusações.
  - Se você não documentar os problemas, não poderá corrigi-los.
    - Garanta que a análise pós-incidente seja isenta de acusações para que você possa ser imparcial em relação às ações corretivas propostas e promover uma autoavaliação e uma colaboração justas às equipes de aplicativos.
- Use um processo para determinar fatores contribuintes. Tenha um processo para identificar e documentar os fatores que contribuem para um evento para que você possa desenvolver mitigações a fim de limitar ou impedir a recorrência e elaborar procedimentos para respostas rápidas e eficazes. Comunique os fatores contribuintes conforme apropriado, de acordo com o público-alvo.
  - O que é análise de log?

Recursos

Documentos relacionados:

- O que é análise de log?
- Por que você deve desenvolver uma correção de erro (COE)

REL12-BP03 Testar os requisitos funcionais

Use técnicas como testes de unidade e de integração que validam a funcionalidade necessária.

Você obtém os melhores resultados quando esses testes são executados automaticamente como parte das ações de compilação e implantação. Por exemplo, usando o AWS CodePipeline, os desenvolvedores confirmam alterações em um repositório de origem onde o CodePipeline detecta automaticamente as alterações. Essas alterações são criadas e os testes são executados. Após a conclusão dos testes, o código criado é implantado nos servidores de preparação para testes. No servidor de preparação, o CodePipeline executa mais testes, como testes de integração ou carga. Após a conclusão bem-sucedida desses testes, o CodePipeline implanta o código testado e aprovado nas instâncias de produção.

Além disso, a experiência mostra que o teste de transações sintéticas (também conhecido como teste canário, que não deve ser confundido com as implantações canário) que pode executar e simular o comportamento do cliente está entre os processos de teste mais importantes. Execute esses testes constantemente nos endpoints da carga de trabalho de diversos locais remotos. O Amazon CloudWatch Synthetics permite criar canários para monitorar seus endpoints e APIs.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Teste os requisitos funcionais. Esse procedimento inclui testes de unidade e de integração que validam a funcionalidade necessária.
  - Usar o CodePipeline com o AWS CodeBuild para testar código e executar builds
  - O AWS CodePipeline adiciona compatibilidade para testes de unidade e de integração personalizada com o AWS CodeBuild
  - Entrega contínua e integração contínua
  - Uso de canários (Amazon CloudWatch Synthetics)
  - Automação de teste de software

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar na implementação de um pipeline de integração contínua
- O AWS CodePipeline adiciona compatibilidade para testes de unidade e de integração personalizada com o AWS CodeBuild
- AWS Marketplace: produtos que podem ser usados para integração contínua
- Entrega contínua e integração contínua
- Automação de teste de software
- Usar o CodePipeline com o AWS CodeBuild para testar código e executar builds
- Uso de canários (Amazon CloudWatch Synthetics)

REL12-BP04 Testar os requisitos de escalabilidade e performance

Use técnicas como o teste de carga para validar se a workload atende aos requisitos de escalabilidade e performance.

Na nuvem, você pode criar um ambiente de teste em escala de produção sob demanda para sua carga de trabalho. Se você executar esses testes na infraestrutura reduzida, deverá escalar os resultados observados para o que você acha que acontecerá na produção. Os testes de carga e performance também podem ser feitos na produção se você tiver cuidado para não afetar os usuários reais e marcar seus dados de teste para que eles não se sintam com dados reais do usuário e estatísticas de uso corrompidas ou relatórios de produção.

Com os testes, certifique-se de que seus recursos básicos, configurações de escalabilidade, cotas de serviço e design de resiliência operem conforme o esperado sob carga.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Teste os requisitos de escalabilidade e performance. Execute o teste de carga para validar se a carga de trabalho atende aos requisitos de escalabilidade e performance.
  - Distributed Load Testing on AWS: simulate thousands of connected users (Teste de carga distribuída na AWS: simular milhares de usuários conectados)
  - Apache JMeter
    - Implante seu aplicativo em um ambiente idêntico ao seu ambiente de produção e execute um teste de carga.
      - Use os conceitos de infraestrutura como código para criar um ambiente que seja o mais semelhante possível ao seu ambiente de produção.

Recursos

Documentos relacionados:

- Distributed Load Testing on AWS: simulate thousands of connected users (Teste de carga distribuída na AWS: simular milhares de usuários conectados)
- Apache JMeter

REL12-BP05 Testar a resiliência por meio da engenharia do caos

Execute testes de caos regularmente em ambientes que estão em produção, ou muito próximos de entrarem em produção, para entender como seu sistema responde a condições adversas.

Resultado desejado:

A resiliência da workload é regularmente verificada por meio da aplicação de engenharia de caos na forma de testes de injeção de falha ou injeção de carga inesperada, além de testes de resiliência que validam o comportamento conhecido esperado da workload durante um evento. Combine engenharia de caos e testes de resiliência para ter confiança de que sua workload poderá sobreviver à falha de componentes e se recuperar de interferências inesperadas com pouco ou nenhum impacto.

Antipadrões comuns:

- Projetar para resiliência, mas não verificar como a workload funciona como um todo quando ocorrem falhas.
- Nunca realizar testes sob condições reais e carga esperada.
- Não tratar seus testes como código nem mantê-los ao longo do ciclo de desenvolvimento.
- Não realizar testes de caos tanto como parte do pipeline de CI/CD quanto fora das implantações.
- Negar o uso de análises pós-incidentes passadas ao determinar quais falhas usar para realizar testes.

Benefícios do estabelecimento desta prática recomendada: A injeção de falhas para verificar a resiliência de uma workload permite que você obtenha confiança de que os procedimentos de recuperação de seu design resiliente vão funcionar em caso de falha real.

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

Orientação de implementação

A engenharia de caos proporciona à sua equipe os recursos para injetar continuamente interferências (simulações) reais de maneira controlada no provedor de serviço, na infraestrutura, na workload e no componente, com pouco ou nenhum impacto para os clientes. Permite que as equipes aprendam com as falhas e observem, mensurem e aumentem a resiliência das workloads, além de validar o acionamento de alertas e a notificação das equipes em caso de evento.

Quando realizada continuamente, a engenharia de caos pode destacar deficiências nas workloads que, se não respondidas, podem afetar negativamente a disponibilidade e a operação.

> **Note**
>
> A engenharia do caos é a disciplina de experimentar um sistema distribuído para aumentar a confiança na capacidade do sistema de resistir a condições turbulentas na produção. – Princípios da engenharia do caos

Se um sistema é capaz de suportar essas interferências, os testes de caos devem ser mantidos como testes de regressão automatizados. Dessa forma, os testes de caos devem ser realizados como parte do ciclo de vida de desenvolvimento dos sistemas (SDLC) e como parte do pipeline de CI/CD.

Para garantir que sua workload pode sobreviver à falha de componentes, injete eventos reais como parte dos testes. Por exemplo, realize testes com perda de instâncias do Amazon EC2 ou failover da instância de banco de dados primária do Amazon RDS e verifique se a workload não é afetada (ou apenas minimamente afetada). Use uma combinação de falhas de componentes para simular eventos que podem ser causados por uma interferência em uma zona de disponibilidade.

Para falhas no nível da aplicação (como travamentos), você pode começar com fatores de estresse, como exaustão de memória e CPU.

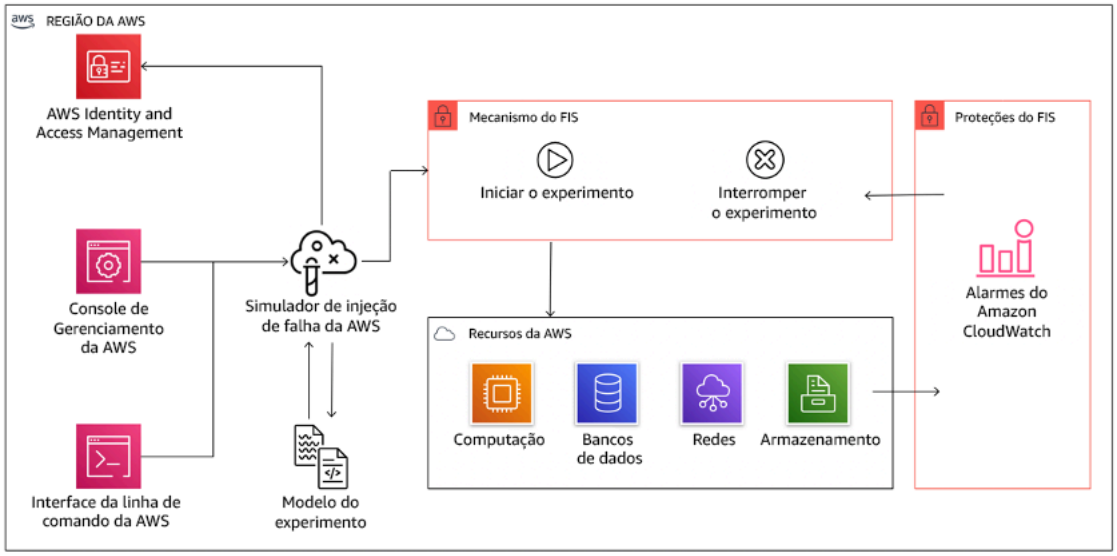
Para validar mecanismos de fallback ou failover para dependências externas devido a interferências intermitentes na rede, os componentes devem simular esse tipo de evento bloqueando o acesso aos provedores externos durante um período especificado, que pode variar de segundos a horas.

Outros modos de degradação podem levar a uma redução nas funcionalidades e a respostas lentas, muitas vezes levando a uma interrupção dos serviços. Essa degradação costuma resultar de um aumento na latência de serviços críticos e comunicação de rede não confiável (pacotes abandonados). Testes com essas falhas, incluindo efeitos de rede como latência, mensagens perdidas e falhas de DNS, podem incluir a incapacidade de resolver um nome, alcançar o serviço de DNS ou estabelecer conexões com serviços dependentes.

Ferramentas de engenharia de caos:

o AWS Fault Injection Service (AWS FIS) é um serviço totalmente gerenciado para a execução de experimentos de injeção de falha que podem ser usados como parte do pipeline de CD, ou fora do pipeline. O AWS FIS é uma boa opção para ser usado durante dias de jogo de engenharia de caos. Oferece suporte à introdução simultânea de falhas em diferentes tipos de recursos, incluindo Amazon EC2, Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service

(Amazon EKS) e Amazon RDS. Essas falhas incluem encerramento de recursos, failovers forçados, esgotamento de CPU ou memória, controle de utilização, latência e perda de pacotes. Por ser integrado a alarmes do Amazon CloudWatch, você pode definir condições de parada como barreiras de proteção para reverter um teste se causar impacto inesperado.



O AWS Fault Injection Service se integra a recursos da AWS para permitir a execução de experimentos de injeção de falha para as workloads.

Existem também várias opções de terceiros para experimentos de injeção de falhas. Elas incluem ferramentas de código aberto, como o Chaos Toolkit, Chaos Meshe aos Litmus Chaos, bem como opções comerciais como o Gremlin. Para expandir o escopo de falhas que podem ser injetadas na AWS, o AWS FIS integra-se ao Chaos Mesh e Litmus Chaos, possibilitando que você coordene fluxos de trabalho de injeção de falhas entre várias ferramentas. Por exemplo, você pode executar um teste de estresse na CPU de um pod usando falhas do Chaos Mesh ou Litmus enquanto encerra uma porcentagem selecionada aleatoriamente de nós de cluster usando ações de falha do AWS FIS.

Etapas da implementação

- Determine quais falhas usar para os testes.

  Avalie o design de sua workload quanto à resiliência. Tais designs (criados usando as práticas recomendadas do Well-Architected Framework) consideram riscos baseados em dependências críticas, eventos passados, problemas conhecidos e requisitos de conformidade. Liste cada elemento do design destinado a manter a resiliência e as falhas que foi projetado para mitigar. Para obter mais informações sobre a criação dessas listas, consulte o artigo técnico Análise de prontidão operacional que orienta você sobre como criar um processo para impedir a recorrência

de incidentes passados. O processo de modos de falhas e análises de efeitos (FMEA) proporciona um framework para realização de análise de falhas em nível de componente e como elas afetam a workload. O FMEA foi descrito em mais detalhes por Adrian Cockcroft em Failure Modes and Continuous Resilience (Modos de falhas e resiliência contínua).

- Atribua uma prioridade a cada falha.

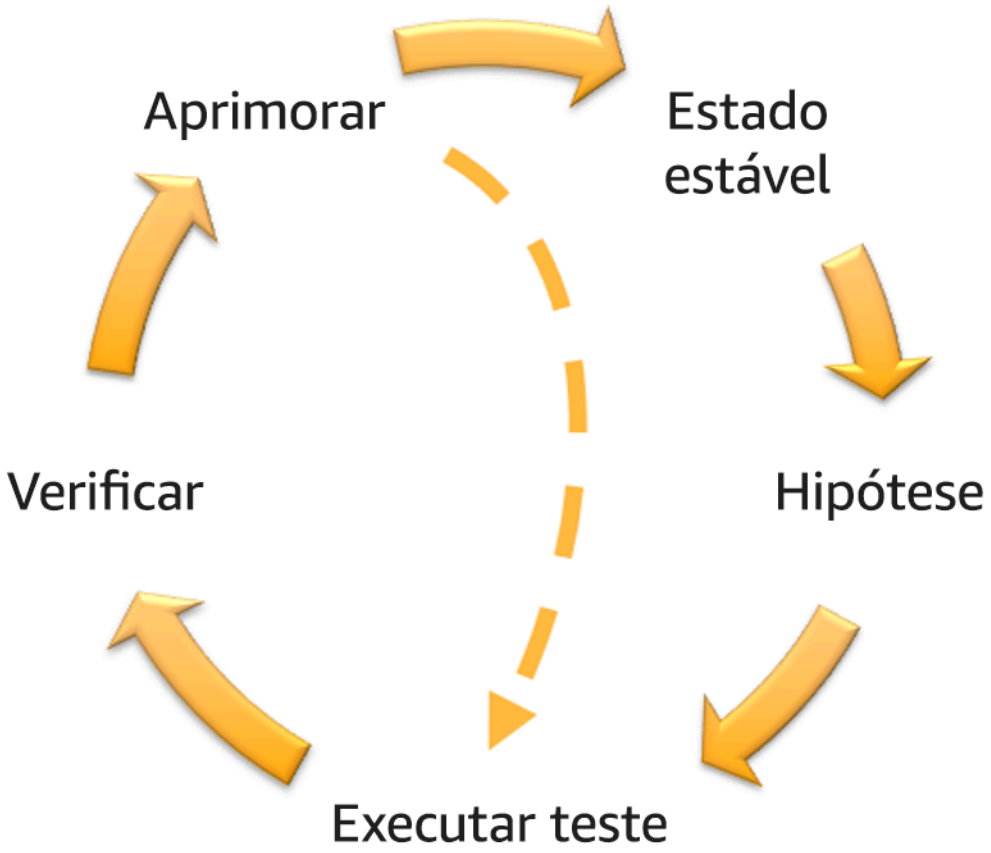
  Comece com uma categorização bruta, como alta, média e baixa. Para avaliar a prioridade, considere a frequência da falha e o impacto da falha na workload total.

  Ao considerar a frequência de determinada falha, analise os dados passados para essa workload sempre que disponíveis. Caso contrário, use os dados de outras workloads executadas em ambientes semelhantes.

  Ao considerar o impacto de determinada falha, em geral, quanto maior o escopo da falha, maior o impacto. Considere também o design e a finalidade da workload. Por exemplo, a capacidade de acessar os datastores de origem é essencial para uma workload que executa análise e transformação de dados. Nesse caso, priorize testes de falhas de acesso, além de acesso controlado e inserção de latência.

  Análises pós-incidente são boas fontes de dados para entender a frequência e o impacto dos modos de falha.

  Use a prioridade atribuída para determinar quais falhas escolher para testar primeiro e a sequência para desenvolver novos testes de injeção de falhas.
- Para cada teste realizado, siga o flywheel de engenharia de caos e resiliência contínua.



Flywheel de engenharia de caos e resiliência contínua, usando o método científico por Adrian Hornsby.

- Defina o estado estável como uma saída mensurável de uma workload que indica comportamento normal.

  Sua workload apresentará estado estável se estiver operando de maneira confiável e conforme o esperado. Portanto, valide a integridade da workload antes de definir o estado estável. O estado estável nem sempre significa que não há nenhum impacto à workload quando ocorre uma falha, já que determinada porcentagem de falhas pode estar dentro de limites aceitáveis. O estado estável é a linha de base que você vai observar durante o teste, o que vai destacar anomalias se a hipótese definida na próxima etapa não sair conforme o esperado.

  Por exemplo, um estado estável de um sistema de pagamentos pode ser definido como o processamento de 300 TPS com taxa de sucesso de 99% e tempo de ida e volta de 500 ms.

- Formule uma hipótese sobre como a workload vai reagir à falha.

Uma boa hipótese se baseia em como se espera que a workload mitigue a falha para manter o estado estável. A hipótese afirma que para determinado tipo de falha, o sistema ou a workload vai permanecer em estado estável, pois a workload foi projetada com mitigações específicas. O tipo específico de falhas e mitigações deve ser especificado na hipótese.

O modelo a seguir pode ser usado para a hipótese (mas outras palavras também são aceitáveis):

> **Note**
>
> Se *falha específica* ocorrer, a workload *nome da workload* vai *descrever os controles de mitigação* para manter *impacto da métrica de negócios ou técnica*.

Por exemplo:

- Se 20% dos nós no grupo de nós do Amazon EKS forem desativados, a API Transaction Create continuará atendendo ao 99.º percentil das solicitações em menos de 100 ms (estado estável). Os nós do Amazon EKS vão se recuperar em cinco minutos e os pods serão agendados e processarão o tráfego oito minutos depois do início do experimento. Os alertas serão acionados em três minutos.
- Se ocorrer uma única falha de instância do Amazon EC2, a verificação de integridade do Elastic Load Balancing do sistema de ordem vai fazer com que o Elastic Load Balancing envie solicitações apenas para as instâncias íntegras restantes, enquanto o Amazon EC2 Auto Scaling substitui a instância com falha, mantendo um aumento inferior a 0,01% na quantidade de erros no servidor (5xx) (estado estável).
- Se a instância de banco de dados primária do Amazon RDS falhar, a workload de coleta de dados da cadeia de suprimentos vai entrar em failover e se conectará à instância de banco de dados de espera do Amazon RDS para manter menos de um minuto de erros de leitura ou gravação de banco de dados (estado estável).

- Execute o teste injetando a falha.

Um teste deve, por padrão, ser seguro contra falhas e tolerado pela workload. Se você sabe que a workload vai falhar, não execute o teste. A engenharia de caos deve ser usada para encontrar incertezas conhecidas ou desconhecidas. Incertezas conhecidas são coisas que você conhece, mas não entende totalmente, enquanto incertezas desconhecidas são coisas que você não

conhece nem entende totalmente. Realizar testes em uma workload que você sabe que está quebrada não oferecerá novos insights. Seu teste deve ser cuidadosamente planejado, ter um escopo claro do impacto e fornecer um mecanismo de reversão que possa ser aplicado em caso de turbulência inesperada. Se sua devida diligência mostrar que a workload sobreviverá ao teste, prossiga com o teste. Há diversas opções para injetar as falhas. Para workloads na AWS, AWS FIS oferece diversas simulações de falhas predefinidas chamadas de ações. Você também pode definir ações personalizadas que são executadas no AWS FIS usando documentos do AWS Systems Manager.

Nós desencorajamos o uso de scripts personalizados para testes de caos, a menos que os scripts tenham os recursos para entender o estado atual da workload, sejam capazes de emitir logs e ofereçam mecanismos para rollbacks e condições de parada sempre que possível.

Um conjunto de ferramentas ou framework eficaz que ofereça suporte à engenharia de caos deve monitorar o estado atual de um experimento, emitir logs e fornecer mecanismos de rollback para oferecer suporte à execução controlada de um teste. Comece com um serviço estabelecido, como o AWS FIS, que permita que você realize testes com um escopo claramente definido e mecanismos de segurança que reverterão o teste se ele introduzir turbulência inesperada. Para conhecer uma ampla variedade de testes que usam o AWS FIS, consulte também o laboratório Aplicações resilientes e bem-arquitetadas com engenharia de caos. Além disso, o AWS Resilience Hub vai analisar sua workload e criar testes que você pode escolher para implementação e execução no AWS FIS.

> **Note**
>
> Para cada teste, entenda claramente o escopo e seu impacto. Recomendamos que as falhas sejam simuladas primeiro em um ambiente que não seja de produção, antes de serem executadas em produção.

Os testes devem ser executados em produção sob carga real usando implantações canário que ativam implantações de controle e experimentais no sistema, sempre que viável. A realização de testes durante horários fora de pico é uma boa prática para mitigar o impacto potencial durante o primeiro teste em produção. Além disso, se o uso de tráfego real de clientes for algo muito arriscado, você poderá executar testes usando tráfego sintético na infraestrutura de produção em implantações de controle e experimentais. Quando não for possível usar a produção, realize os testes em ambientes de pré-produção que sejam o mais parecido possível com produção.

Estabeleça e monitore barreiras de proteção para garantir que o teste não afete o tráfego de produção ou outros sistemas além dos limites aceitáveis. Estabeleça condições de parada para interromper um teste se ele atingir um limite definido de uma métrica de barreira de proteção. Isso deve incluir as métricas de estado estável da workload, bem como a métrica em relação aos componentes em que você está injetando a falha. A monitor sintético (também conhecido como canário de usuário) é uma métrica que geralmente deve ser incluída como proxy de usuário. Condições de parada do AWS FIS são compatíveis como parte do modelo de teste, permitindo até cinco condições de parada por modelo.

Um dos princípios de caos é minimizar o escopo do teste e seu impacto:

embora deva existir uma provisão para algum impacto negativo de curto prazo, é responsabilidade e obrigação do engenheiro de caos garantir que as perdas dos testes sejam minimizadas e contidas.

Um método para verificar o escopo e o impacto potencial é realizar o teste primeiro em um ambiente que não seja de produção, verificando se os limites para as condições de parada são ativados conforme o esperado durante o teste e se há observabilidade em vigor para identificar uma exceção, em vez de testar diretamente em produção.

Ao executar testes de injeção de falhas, verifique se todas as partes responsáveis estão bem informadas. Comunique-se com as equipes adequadas, como equipes de operações, equipes de confiabilidade do serviço e atendimento ao cliente, para avisá-las sobre quando os testes serão realizados e o que esperar. Ofereça a essas equipes ferramentas de comunicação para que informem os responsáveis pela execução do teste caso percebam algum efeito adverso.

Você deve restaurar a workload e seus sistemas subjacentes de volta para o estado íntegro original. Normalmente, o design resiliente da workload vai se autorrestaurar. No entanto, alguns designs de falhas ou testes malsucedidos podem deixar a workload em um estado de falha inesperado. Ao final do teste, você deverá estar ciente disso e restaurar a workload e os sistemas. Com o AWS FIS, você pode definir uma configuração de reversão (também chamada de ação posterior) nos parâmetros de ação. Uma ação posterior restaura o destino para o estado em que estava antes da execução da ação. Independentemente de serem automatizadas (como as que usam o AWS FIS) ou manuais, essas ações posteriores devem fazer parte de um playbook que descreve como detectar e lidar com falhas.

- Verifique a hipótese.

Princípios da engenharia do caos oferecem a seguinte orientação sobre como verificar o estado estável de sua workload:

Concentre-se na saída mensurável de um sistema, em vez de atributos internos do sistema. As medições dessa saída durante um curto período constituem um proxy do estado estável do sistema. A throughput total do sistema, as taxas de erros e os percentis de latência podem ser métricas de interesse que representam o comportamento do estado estável. Ao focar em padrões de comportamento sistêmicos durante os testes, a engenharia de caos verifica se o sistema de fato funciona em vez de tentar validar como ele funciona.

Nos dois exemplos anteriores, nós incluímos as métricas de estado estável de menos de 0,01% de aumento na quantidade de erros no servidor (5xx) e menos de um minuto de erros de leitura ou gravação de banco de dados.

Os erros 5xx são uma boa métrica, pois são consequência do modo de falha que um cliente da workload vai vivenciar diretamente. A medição dos erros do banco de dados é boa como consequência direta da falha, mas também deve ser complementada com uma medição de impacto para o cliente, como solicitações malsucedidas ou erros apresentados ao cliente. Além disso, inclua um monitor sintético (também conhecido como canário de usuário) em todas as APIs ou URIs acessadas pelo cliente da workload.

- Melhore o design da workload para agregar resiliência.

  Se o estado estável não tiver sido mantido, investigue como o design da workload pode ser melhorado para mitigar a falha, aplicando as práticas recomendadas do pilar Confiabilidade do AWS Well-Architected. Orientação e recursos adicionais podem ser encontrados na AWS Builder's Library, que contém artigos sobre como melhorar as verificações de integridade ou implantar repetições sem recuo no código de sua aplicação, entre outros.

  Depois de implementar essas mudanças, execute o teste novamente (mostrado pela linha pontilhada no flywheel de engenharia de caos) para determinar a eficácia. Se a etapa de verificação indicar que a hipótese é verdadeira, a workload estará em estado estável e o ciclo continuará.

- Execute testes regularmente.

  Um teste de caos é um ciclo, e os testes devem ser realizados regularmente como parte da engenharia de caos. Depois que uma workload cumprir a hipótese do teste, o teste deverá ser automatizado para ser executado continuamente como parte de regressão do pipeline de CI/CD.

Para saber como fazer isso, consulte este blog sobre como executar testes do AWS FIS usando o AWS CodePipeline. Este laboratório sobre testes recorrentes do AWS FIS em um pipeline de CI/CD permite que você trabalhe de maneira prática.

Os testes de injeção de falhas também fazem parte dos dias de jogo (consulte REL12-BP06 Realizar dias de jogo regularmente). Os dias de jogo simulam uma falha ou um evento para verificar sistemas, processos e respostas das equipes. O objetivo é realmente executar as ações que a equipe executaria como se um evento excepcional acontecesse.

- Capture e armazene os resultados do teste.

  Os resultados da injeção de falhas devem ser capturados e persistidos. Inclua todos os dados necessários (como tempo, workload e condições) para poder analisar os resultados e as tendências do teste posteriormente. Exemplos de resultados podem incluir capturas de tela de painéis, despejos em CSV do banco de dados da métrica ou um registro manual dos eventos e das observações do teste. O registro do teste em log com o AWS FIS pode fazer parte dessa captura de dados.

Recursos

Práticas recomendadas relacionadas:

- REL08-BP03 Integrar testes de resiliência como parte da sua implantação
- REL13-BP03 Testar a implementação de recuperação de desastres para validá-la

Documentos relacionados:

- O que é o AWS Fault Injection Service?
- O que é o AWS Resilience Hub?
- Princípios da engenharia do caos
- Engenharia de caos: planejando seu primeiro teste
- Engenharia de resiliência: aprendendo a aceitar falhas
- Histórias sobre engenharia de caos
- Evitar fallback em sistemas distribuídos
- Implantação canário para testes de caos

Vídeos relacionados:

- AWS re:Invent 2020: Testing resiliency using chaos engineering (ARC316) (AWS re:Invent 2020: teste de resiliência usando engenharia de caos)
- AWS re:Invent 2019: Improving resiliency with chaos engineering (DOP309-R1) (AWS re:Invent 2019: melhoria da resiliência com engenharia de caos)
- AWS re:Invent 2019: Performing chaos engineering in a serverless world (CMY301) (AWS re:Invent 2019: execução da engenharia de caos em um universo de tecnologia sem servidor)

Exemplos relacionados:

- Laboratório do Well-Architected: nível 300: testes de resiliência do Amazon EC2, Amazon RDS e Amazon S3
- Laboratório Engenharia de caos na AWS
- Laboratório Aplicações resilientes e bem-arquitetadas com engenharia de caos
- Laboratório Caos em tecnologia sem servidor
- Laboratório Mensurar e aumentar a resiliência de sua aplicação com o AWS Resilience Hub

Ferramentas relacionadas:

- AWS Fault Injection Service
- AWS Marketplace: plataforma de engenharia de caos Gremlin
- Chaos Toolkit
- Chaos Mesh
- Litmus

REL12-BP06 Realizar dias de jogo regularmente

Use os dias de jogo para praticar regularmente seus procedimentos de resposta a eventos e falhas o mais próximo possível da produção (inclusive em ambientes de produção) e com as pessoas que estarão envolvidas nos cenários de falha reais. Os dias de jogo aplicam medidas para garantir que os eventos de produção não afetem os usuários.

Os dias de jogo simulam uma falha ou evento para testar sistemas, processos e respostas das equipes. O objetivo é realmente executar as ações que a equipe executaria como se um evento

excepcional acontecesse. Isso ajudará a compreender onde as melhorias podem ser feitas e pode ajudar a desenvolver experiência organizacional ao lidar com eventos. Eles devem ser realizados regularmente para que a equipe desenvolva memória muscular sobre como responder.

Depois que o projeto de resiliência estiver em vigor e tiver sido testado em ambientes que não sejam de produção, um dia de jogo será a maneira de garantir que tudo funcione conforme o planejado na produção. Um dia de jogo, especialmente o primeiro, é uma atividade de "todos os funcionários" em que engenheiros e operações são informados quando isso acontecerá e o que ocorrerá. Há runbooks disponíveis. Os eventos simulados são executados, incluindo possíveis eventos de falha, nos sistemas de produção da maneira prescrita, e o impacto é avaliado. Se todos os sistemas operarem conforme projetado, a detecção e a recuperação automática ocorrerão com pouco ou nenhum impacto. No entanto, se houver impacto negativo, o teste será revertido e os problemas da workload serão corrigidos manualmente, se necessário (usando o runbook). Como os dias de jogos ocorrem na produção, todas as precauções devem ser tomadas para garantir que não haja impacto na disponibilidade dos clientes.

Antipadrões comuns:

- Documentar seus procedimentos, mas nunca os praticar.
- Não incluir os tomadores de decisão de negócios nos exercícios de teste.

Benefícios do estabelecimento desta prática recomendada: A realização frequente dos dias de jogo garante que toda a equipe siga e valide as políticas e os procedimentos apropriados quando ocorrer um incidente real.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Programe os dias de jogo para praticar regularmente os runbooks e os manuais. Os dias de jogo devem incluir todas as pessoas envolvidas em um evento de produção: proprietário da empresa, equipe de desenvolvimento, equipe operacional e equipes de resposta a incidentes.
  - Execute os testes de carga ou de performance e, em seguida, execute a injeção de falha.
  - Procure por anomalias nos runbooks e oportunidades de praticar os playbooks.
    - Se você se desviar dos runbooks, refine-os ou corrija o comportamento. Se você praticar o playbook, identifique o runbook que deveria ter sido usado ou crie um novo.

Recursos

Documentos relacionados:

- O que é o AWS GameDay?

Vídeos relacionados:

- AWS re:Invent 2019: Improving resiliency with chaos engineering (DOP309-R1)

Exemplos relacionados:

- Laboratórios do AWS Well-Architected: testes de resiliência

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. RTO e RPO são os seus objetivos para restauração de sua workload. 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. A probabilidade de interrupção e o custo de recuperação também são fatores principais que ajudam a determinar o valor empresarial de fornecer a recuperação de desastres para uma workload.

Práticas recomendadas

- REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados
- REL13-BP02 Usar estratégias de recuperação definidas para atender aos objetivos de recuperação
- REL13-BP03 Testar a implementação de recuperação de desastres para validá-la
- REL13-BP04 Gerenciar o desvio de configuração para o local ou a região de DR
- REL13-BP05 Automatizar a recuperação

REL13-BP01 Definir 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).

Recovery Time Objective (RTO – Objetivo do tempo de recuperação) é o atraso máximo aceitável entre a interrupção do serviço e sua restauração. Isso determina o que é considerado uma janela de tempo aceitável quando o serviço está indisponível.

Recovery Point Objective (RPO – Objetivo do ponto de recuperação) é o tempo máximo aceitável desde o último ponto de recuperação de dados. Isso determina o que é considerado uma perda aceitável de dados entre o último ponto de recuperação e a interrupção do serviço.

Os valores de RTO e RPO são considerações importantes ao selecionar uma estratégia de recuperação de desastres (DR) apropriada para a workload. Esses objetivos são determinados pelo negócio e, em seguida, usados pelas equipes técnicas para selecionar e implementar uma estratégia de DR.

Resultado desejado:

Cada workload tem um RTO e um RPO atribuídos, definidos com base no impacto empresarial. A workload é atribuída a uma camada predefinida com um RTO e um RPO associados, estabelecendo a disponibilidade do serviço e a perda aceitável de dados. Se isso não for possível, poderá ser atribuído sob medida por workload com a intenção de criar camadas posteriormente. O RTO e o RPO são usados como uma das principais considerações para a seleção da implementação de uma estratégia de recuperação de desastres para a workload. São considerações adicionais na escolha de uma estratégia de DR as restrições de custo, as dependências da workload e os requisitos operacionais.

Para o RTO, compreenda o impacto com base na duração de uma interrupção. É linear ou há implicações não lineares? (Por exemplo, após quatro horas, você desliga uma linha de produção até o início do próximo turno).

Uma matriz de recuperação de desastres, como a seguinte, pode ajudar você a compreender como a criticidade da workload se relaciona com os objetivos de recuperação. (Observe que os valores reais dos eixos X e Y devem ser personalizados de acordo com as necessidades da sua organização).

Matriz de recuperação de desastres						
		Objetivo do ponto de recuperação				
		< 1 minuto	< 1 hora	< 6 horas	< 1 dia	+ 1 dia
Objetivo do tempo de recuperação	< 10 minutos	Crítica	Crítica	Alto	Médio	Médio
	< 2 horas	Crítica	Alto	Médio	Médio	Baixo
	< 8 horas	Alto	Médio	Médio	Baixo	Baixo
	< 24 horas	Médio	Médio	Baixo	Baixo	Baixo
	+ de 24 horas	Médio	Baixo	Baixo	Baixo	Baixo

Figura 16: Matriz de recuperação de desastres

Antipadrões comuns:

- Objetivos de recuperação não definidos.
- Seleção de objetivos de recuperação arbitrários.
- Seleção de objetivos de recuperação que são muito permissivos e não atendem aos objetivos de negócios.
- Não compreender o impacto do tempo de inatividade e da perda de dados.
- Seleção de objetivos de recuperação irreais, como nenhum tempo para recuperação e nenhuma perda de dados, que podem não ser alcançáveis para a configuração da workload.
- Seleção de objetivos de recuperação mais rigorosos do que os objetivos de negócios reais. Isso força implementações de DR mais caras e complicadas do que as necessidades da workload.
- Seleção de objetivos de recuperação incompatíveis com os da workload dependente.
- Os objetivos de recuperação não consideram os requisitos regulamentares de conformidade.
- RTO e RPO definidos para uma workload, mas nunca testados.

Benefícios do estabelecimento dessa prática recomendada: Os objetivos de recuperação referentes a tempo e perda de dados são necessários para orientar a implementação da DR.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

Para a workload, você deve compreender o impacto do tempo de inatividade e da perda de dados em seus negócios. O impacto geralmente aumenta com maior tempo de inatividade ou perda de dados, mas a forma desse crescimento pode diferir com base no tipo de workload. Por exemplo, pode ser que você consiga tolerar o tempo de inatividade por até uma hora com pouco impacto, mas depois disso o impacto aumenta rapidamente. O impacto nos negócios se manifesta de diversas formas, incluindo custo monetário (como perda de receita), confiança do cliente (e impacto na reputação), problemas operacionais (como folha de pagamento ausente ou diminuição na produtividade) e risco regulatório. Use as etapas a seguir para compreender esses impactos e defina o RTO e o RPO para sua workload.

Etapas da implementação

1. Determine as partes interessadas do negócio para a workload e interaja com eles para implementar essas etapas. Os objetivos de recuperação para uma workload são uma decisão de negócios. As equipes técnicas trabalham com as partes interessadas do negócio para usar esses objetivos para selecionar uma estratégia de DR.

   > **Note**
   >
   > Para as etapas 2 e 3, você pode usar o the section called “Planilha de implementação”.

2. Reúna as informações necessárias para tomar uma decisão respondendo às perguntas abaixo.
3. Você tem categorias ou níveis de criticidade para o impacto da workload na sua organização?
   a. Se sim, atribua esta workload a uma categoria
   b. Se não, estabeleça estas categorias. Crie cinco ou menos categorias e refine o intervalo do seu objetivo de tempo de recuperação para cada uma delas. Os exemplos de categorias incluem: crítica, alta, média, baixa. Para entender como uma workloads é mapeada para uma categoria, considere se ela é de missão crítica, importante para os negócios ou não comercial.
   c. Defina o RTO e o RPO da workload com base na categoria. Sempre escolha uma categoria mais restrita (RTO e RPO mais baixos) do que os valores brutos calculados no começo desta etapa. Se isso resultar em uma mudança de valor inadequadamente grande, considere a criação de uma nova categoria.
4. Com base nessas respostas, atribua valores de RTO e RPO à workload. Isso pode ser feito diretamente ou atribuindo a workload a uma camada de serviço predefinida.

5. Documente o plano de recuperação de desastres (DRP) para esta workload, que faz parte do plano de continuidade de negócios (BCP) da sua organização, em um local acessível à equipe de workload e às partes interessadas
   a. Registre o RTO, o RPO e as informações usadas para determinar esses valores. Inclua a estratégia usada para avaliar o impacto da workload nos negócios.
   b. Registre outras métricas, além do RTO e do RPO que você está acompanhando ou planeja acompanhar, para os objetivos de recuperação de desastres
   c. Você adicionará detalhes da sua estratégia de DR e runbook a este plano ao criá-los.
6. Ao pesquisar a criticidade da workload em uma matriz, como a da figura 15, você pode começar a estabelecer camadas predefinidas de serviço estabelecidos para sua organização.
7. Após implementar uma estratégia de DR (ou uma prova de conceito para uma estratégia de DR) conforme the section called “REL13-BP02 Usar estratégias de recuperação definidas para atender aos objetivos de recuperação”, teste a estratégia para determinar a capacidade de tempo de recuperação (RTC) e a capacidade de ponto de recuperação (RPC) reais da workload. Se elas não atenderem aos objetivos de recuperação de destino, trabalhe com as partes interessadas do negócio para ajustar esses objetivos ou faça alterações na estratégia de DR para atingir os objetivos de destino.

Perguntas principais

1. Qual é o tempo máximo que a workload pode ficar inativa antes que ocorra um impacto grave nos negócios?
   a. Determine o custo monetário (impacto financeiro direto) para o negócio por minuto se a workload for interrompida.
   b. Considere que o impacto nem sempre é linear. O impacto pode ser limitado no início e aumentar rapidamente após um ponto crítico.
2. Qual é a quantidade máxima de dados que podem ser perdidos antes que ocorra um impacto severo nos negócios?
   a. Considere esse valor para seu armazenamento de dados mais crítico. Identifique a respectiva criticidade para outros armazenamentos de dados.
   b. Os dados de workload podem ser recriados em caso de perda? Se isso for operacionalmente mais fácil do que fazer backup e restauração, escolha o RPO com base na criticidade dos dados de origem usados para recriar os dados da workload.

3. Quais são os objetivos de recuperação e as expectativas de disponibilidade das workloads das quais este depende (downstream) ou as workloads que dependem deste (upstream)?
    a. Escolha objetivos de recuperação que permitam que essa workload atenda aos requisitos das dependências upstream.
    b. Escolha objetivos de recuperação que possam ser alcançados com base nos recursos de recuperação das dependências downstream. Dependências downstream não críticas (aquelas que podem ser “contornadas”) podem ser excluídas. Ou trabalhe com dependências críticas downstream para melhorar os recursos de recuperação quando necessário.

Perguntas adicionais

Considere estas perguntas e como elas podem se aplicar a essa workload:

4. Você tem RTO e RPO diferentes dependendo do tipo de interrupção (região versus AZ etc.)?
5. Existe um momento específico (sazonalidade, eventos de vendas, lançamentos de produtos) em que seu RTO/RPO pode mudar? Se sim, quais são a medição e o limite de tempo diferentes?
6. Quantos clientes serão afetados se a workload for interrompida?
7. Qual será o impacto na reputação se a workload for interrompida?
8. Quais outros impactos operacionais poderão ocorrer se a workload for interrompida? Por exemplo, impacto na produtividade do funcionário se os sistemas de e-mail não estiverem disponíveis ou se os sistemas de folha de pagamento não puderem enviar transações.
9. Como o RTO e o RPO da workload se alinham à estratégia de DR da linha empresarial e organizacional?
10. Há obrigações contratuais internas para a prestação de um serviço? Há penalidades por não cumpri-las?
11. Quais são as restrições regulatórias ou de conformidade com os dados?

Planilha de implementação

Você pode usar esta planilha para as etapas 2 e 3 de implementação. É possível ajustar esta planilha para atender às suas necessidades específicas, como adicionar perguntas.

Etapa 2: Perguntas principais	Aplicável à workload?	RTO da workload	RPO da workload	Ajuste do RTO.	Ajuste do RPO.	Instruções
[1] tempo máximo em que a workload pode ficar inativa						medido com o tempo desde o início da interrupção da recuperação
[2] quantidade máxima de dados que podem ser perdidos						medido com o tempo desde o conjunto de dados bom mais recente restaurável
[3a] dependências upstream						insira os objetivos mais estritos de recuperação upstream
[3b] dependências downstream						insira os objetivos menos estritos de recuperação downstream
[3a] dependências upstream reconciliadas						Se o valor upstream for menor que os valores atuais e o valor downstream for maior,
[3b] dependências downstream reconciliadas						trabalhe com as dependências para fazer a reconciliação e insira os valores reconciliados aqui
[3] dependências						valores menores para atender às dependências upstream ou aumentá-las com base nas capacidades das dependências downstream
Etapa 2: Perguntas adicionais						Indique se a pergunta é aplicável. Se ela não for aplicável, ignore-a
RTO/RPO de base						Carregue os valores de RTO e de RPO acima para baixo, aqui
[4] tipo de interrupção	[ ]S/[ ]N					Insira os objetivos de recuperação para o tipo de evento com os requisitos mais estritos
[5] objetivos baseados em tempo específico	[ ]S/[ ]N					Insira os objetivos de recuperação para momentos com os requisitos mais estritos
[6] clientes interrompidos	[ ]S/[ ]N					Faça um gráfico dos clientes afetados como uma função de tempo de inatividade ou de perda de dados. Use isso para inserir o RTO e o RPO máximos permissíveis com base no impacto no cliente.
[7] impacto na reputação	[ ]S/[ ]N					Trabalhe com a empresa para determinar o RTO e o RPO máximos com base no impacto na reputação
[8] impacto operacional	[ ]S/[ ]N					Insira o RTO e o RPO máximos com base no impacto operacional
[9] alinhamento organizacional	[ ]S/[ ]N					Insira o RTO e o RPO máximos para workloads desse tipo de acordo com as necessidades da LOB e da organização
[10] obrigações contratuais	[ ]S/[ ]N					Insira o RTO e o RPO máximos com base nas obrigações contratuais
[11] conformidade normativa	[ ]S/[ ]N					Insira o RTO e o RPO máximos com base na conformidade normativa aplicável
alvo baseado em questões adicionais						Use o valor mínimo (valor mais estrito) das perguntas 4 a 11 e insira-o aqui
alvo ajustado						Se os objetivos na linha acima não puderem ser acomodados, trabalhe com as partes interessadas para flexibilizar as restrições e insira o novo mínimo aqui
RTO/RPO ajustado						Insira os valores do RPO/RTO de base ou ajuste o alvo, o que for menor
Etapa 3						
Mapear para categoria ou camada predefinida						Ajuste os dois valores para baixo (mais estritos) para que se alinhem com a camada mais próxima definida

Planilha

Nível de esforço para o plano de implementação: Baixo

Recursos

Práticas recomendadas relacionadas:

- the section called “REL09-BP04 Realizar a recuperação periódica dos dados para verificar a integridade e os processos de backup”
- the section called “REL13-BP02 Usar estratégias de recuperação definidas para atender aos objetivos de recuperação”
- the section called “REL13-BP03 Testar a implementação de recuperação de desastres para validá-la”

Documentos relacionados:

- Blog de arquitetura da AWS: série de recuperação de desastres
- Recuperação de desastres de workloads na AWS: recuperação na nuvem (whitepaper da AWS)

- Gerenciamento de políticas de resiliência com o AWS Resilience Hub
- Parceiro do APN: parceiros que podem ajudar com a recuperação de desastres
- AWS Marketplace: produtos que podem ser usados para recuperação de desastres

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
- Recuperação de desastres de workloads na AWS

REL13-BP02 Usar estratégias de recuperação definidas para atender aos objetivos de recuperação

Defina uma estratégia de recuperação de desastres (DR) que atenda os objetivos de recuperação da workload. Escolha uma estratégia como: backup e restauração, standby (ativo-passivo) ou ativo-ativo.

Uma estratégia de DR depende da capacidade de manter a workload em um site de recuperação se seu local primário não puder executar a workload. Os objetivos de recuperação mais comuns são o RTO e o RPO, conforme discutido em REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados.

Uma estratégia de DR em várias zonas de disponibilidade (AZs) em uma única Região da AWS pode fornecer mitigação contra eventos de desastre, como incêndios, inundações e grandes interrupções de energia. Se for um requisito implementar proteção contra um evento improvável que impeça a execução da workload em uma determinada Região da AWS, você poderá optar por uma estratégia de DR que use várias regiões.

Ao arquitetar uma estratégia de DR em várias regiões, você deve escolher uma das seguintes estratégias. Elas estão listadas em ordem crescente de custo e complexidade e em ordem decrescente de RTO e RPO. região de recuperação refere-se a uma Região da AWS diferente da primária usada para a workload.

Figura 17: Estratégias de recuperação de desastres (DR)

- Backup e restauração (RPO em horas, RTO em 24 horas ou menos): faça backup de seus dados e aplicações na região de recuperação. O uso de backups automatizados ou contínuos permitirá a recuperação a um ponto anterior no tempo, podendo reduzir o RPO para até 5 minutos em alguns casos. Em caso de desastre, você implantará a infraestrutura (usando a infraestrutura como código para reduzir o RTO), implantará o código e restaurará os dados salvos para se recuperar de um desastre na região de recuperação.
- Luz piloto (RPO em minutos, RTO em dezenas de minutos): provisione uma cópia da infraestrutura de workload principal na região de recuperação. Replique seus dados na região de recuperação e crie backups deles lá. Os recursos necessários para oferecer suporte à replicação e ao backup, como bancos de dados e armazenamento de objetos, estão sempre ativos. Outros elementos, como servidores de aplicações ou computação com tecnologia sem servidor, não são implantados. Porém, podem ser criados com a configuração e o código da aplicação necessários.
- Modo de espera passivo (RPO em segundos, RTO em minutos): mantenha uma versão reduzida, mas totalmente funcional, da workload sempre em execução na região de recuperação. Os sistemas críticos para os negócios são totalmente duplicados e estão sempre ativados, mas com uma frota reduzida. Os dados são replicados e vivem na região de recuperação. Quando chega o momento da recuperação, o sistema é dimensionado rapidamente para processar a carga de produção. Quanto mais a escala do modo de espera passivo for aumentada verticalmente, menor será a dependência do RTO e do ambiente de gerenciamento. Quando totalmente dimensionado, isso é conhecido como standby a quente.

- Multirregional (multissite) ativo-ativo (RPO próximo a zero, RTO potencialmente zero): a workload é implantada em várias Regiões da AWS e processa ativamente o tráfego delas. Esta estratégia exige que você sincronize os dados entre regiões. Deve-se evitar ou processar possíveis conflitos causados por gravações no mesmo registro em duas réplicas regionais diferentes, o que pode ser complexo. A replicação de dados é útil para a sincronização de dados e protegerá você contra alguns tipos de desastre, mas não contra corrupção ou destruição de dados, a menos que sua solução também inclua opções para recuperação a um ponto anterior no tempo.

> **Note**
>
> Às vezes, a diferença entre luz piloto e modo de espera passivo pode ser difícil de entender. Ambos incluem um ambiente na região de recuperação com cópias dos ativos da região primária. A diferença é que a luz piloto não pode processar solicitações sem primeiro realizar uma ação adicional, enquanto o modo de espera passivo pode processar o tráfego (em níveis de capacidade reduzidos) imediatamente. A luz piloto exigirá que você ative os servidores, possivelmente implante infraestrutura adicional (não essencial) e aumente a escala verticalmente. Já o modo de espera passivo exige apenas que você aumente a escala verticalmente (tudo já está implantado e em execução). Escolha entre elas com base nas suas necessidades de RTO e RPO.

Resultado desejado:

Há uma estratégia de DR definida e implementada para cada workload, permitindo que ela atinja os objetivos de DR. As estratégias de DR entre workloads fazem uso de padrões reutilizáveis (como as estratégias descritas anteriormente).

Antipadrões comuns:

- Implementar procedimentos de recuperação inconsistentes para workloads com objetivos de DR semelhantes.
- Deixar que a estratégia de DR seja implementada ad hoc quando ocorrer um desastre.
- Não tendo nenhum plano para DR.
- Depender das operações do ambiente de gerenciamento durante a recuperação.

Benefícios do estabelecimento dessa prática recomendada:

- O uso de estratégias de recuperação definidas permite que você adote ferramentas comuns e procedimentos de teste.
- O uso de estratégias de recuperação definidas permite o compartilhamento de conhecimento eficiente entre as equipes e uma implementação mais fácil de DR nas workloads que elas possuem.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

- Sem uma estratégia de DR planejada, implementada e testada, é improvável que você atinja os objetivos de recuperação em caso de desastre.

Orientações para a implementação

Para cada uma dessas etapas, veja os detalhes abaixo.

1. Determine uma estratégia de DR que satisfaça os requisitos de recuperação para esta workload.
2. Revise os padrões de como a estratégia de DR selecionada pode ser implementada.
3. Avalie os recursos da workload e qual será sua configuração na região de recuperação antes do failover (durante a operação normal).
4. Determine e implemente como deixar sua região de recuperação pronta para failover quando necessário (durante um evento de desastre).
5. Determine e implemente como redirecionar o tráfego para failover quando necessário (durante um evento de desastre).
6. Projete um plano de como a workload retornará.

Etapas da implementação

1. Determine uma estratégia de DR que satisfaça os requisitos de recuperação para esta workload.

Escolher uma estratégia de DR é uma troca entre reduzir o tempo de inatividade e a perda de dados (RTO e RPO) versus o custo e a complexidade da sua implementação. Você deve evitar implementar uma estratégia mais rigorosa do que necessário, pois isso resulta em custos desnecessários.

Por exemplo, no diagrama a seguir, a empresa determinou seu RTO máximo permitido e o orçamento limite da estratégia de restauração de serviço. Considerando os objetivos do negócio, as

estratégias de DR luz piloto e modo de espera passivo satisfarão tanto o RTO quanto os critérios de custo.

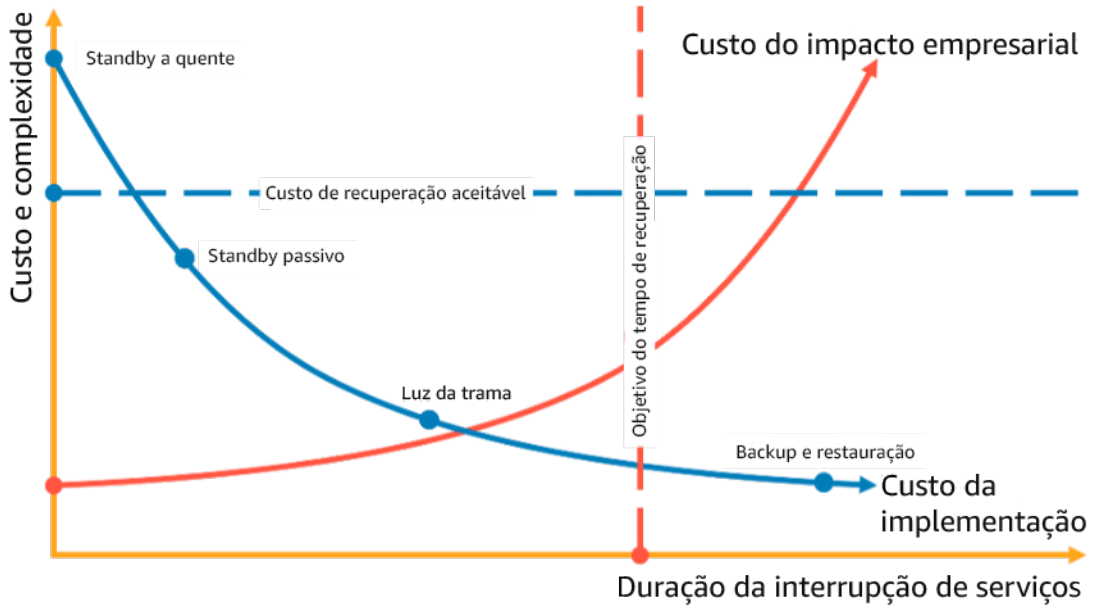


Figura 18: Escolha de uma estratégia de DR com base no RTO e no custo

Para saber mais, consulte Plano de continuidade de negócios (BCP).

2. Revise os padrões de como a estratégia de DR selecionada pode ser implementada.

Esta etapa é para compreender como implementar a estratégia selecionada. As estratégias são explicadas usando as Regiões da AWS como locais primários e de recuperação. No entanto, também é possível optar por usar as zonas de disponibilidade em uma única região como sua estratégia de DR, que faz uso de elementos de várias dessas estratégias.

Nas etapas subsequentes, você aplicará a estratégia à sua workload específica.

Backup e restauração

Backup e restauração é a estratégia menos complexa de implementar. Porém, exigirá mais tempo e esforço para restaurar a workload, levando a RTO e RPO mais altos. É uma boa prática sempre fazer backups dos seus dados e copiá-los para outro local (como outra Região da AWS).

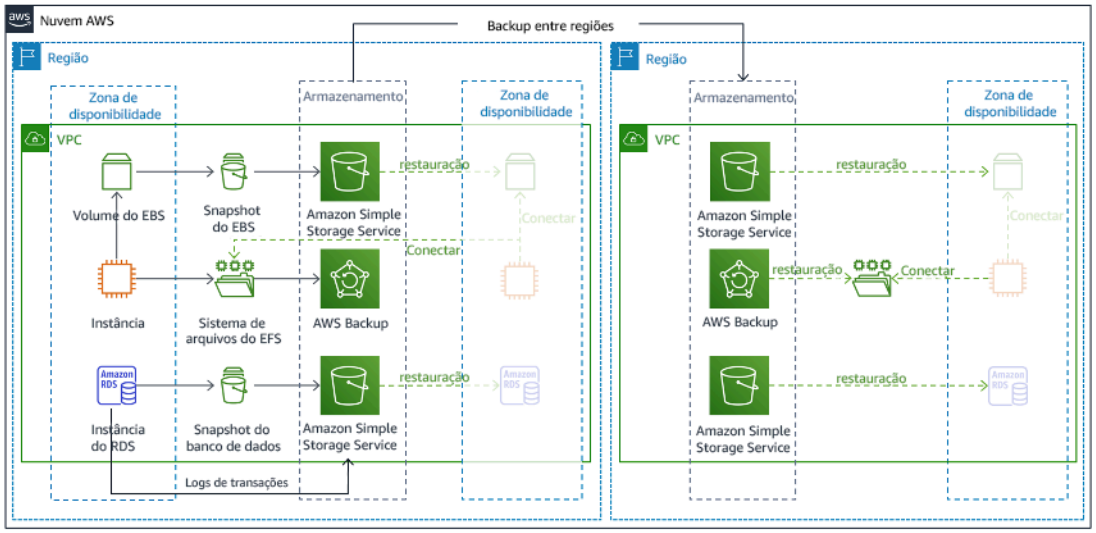


Figura 19: Arquitetura de backup e restauração

Para obter mais detalhes sobre esta estratégia, consulte Arquitetura de recuperação de desastres (DR) na AWS, parte II: backup e restauração com recuperação rápida.

Luz piloto

Com a abordagem de luz piloto , você replica os dados da região primária para a região de recuperação. Os recursos principais usados para a infraestrutura de workload são implantados na região de recuperação. No entanto, recursos adicionais e quaisquer dependências ainda são necessários para tornar isso uma pilha funcional. Por exemplo, na figura 20, nenhuma instância de computação é implantada.

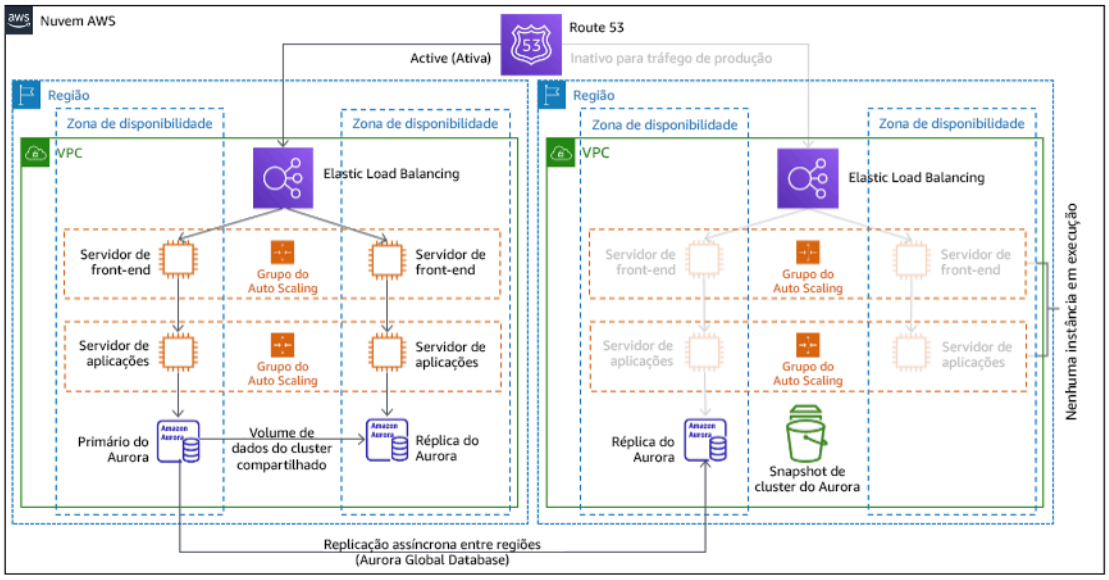


Figura 20: Arquitetura de luz piloto

Para obter mais detalhes sobre esta estratégia, consulte Arquitetura de recuperação de desastres (DR) na AWS, parte III: luz piloto e modo de espera passivo.

Modo de espera passivo

O standby passivo envolve garantir que haja uma cópia com escala reduzida verticalmente, mas totalmente funcional, do seu ambiente de produção em outra região. Essa abordagem estende o conceito de luz piloto e diminui o tempo de recuperação já que a workload está sempre ativa em outra região. A implementação da região de recuperação com capacidade total é conhecido como standby a quente.

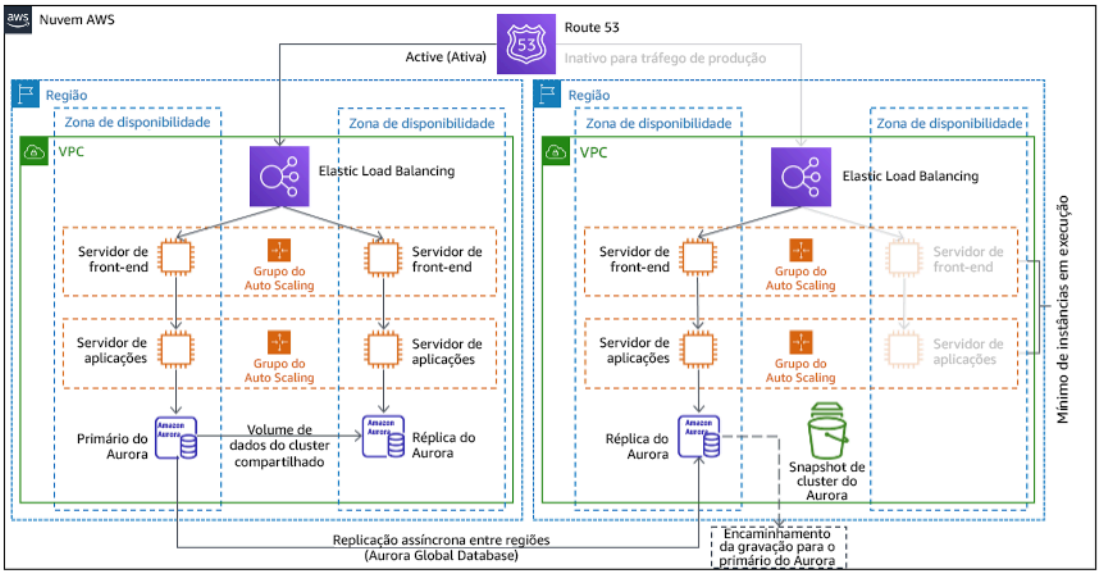


Figura 21: Arquitetura de modo de espera passivo

O uso do modo de espera passivo ou luz piloto requer que a escala dos recursos seja aumentada verticalmente na região de recuperação. Para garantir que a capacidade esteja disponível quando necessário, considere o uso de reservas de capacidade para instâncias do EC2. Se estiver usando o AWS Lambda, a concorrência provisionada poderá garantir que ambientes de execução estejam preparados para responder imediatamente às invocações da sua função.

Para obter mais detalhes sobre esta estratégia, consulte Arquitetura de recuperação de desastres (DR) na AWS, parte III: luz piloto e modo de espera passivo.

Multissite ativo-ativo

Você pode executar sua workload simultaneamente em várias regiões como parte de uma estratégia de multissite ativo-ativo. O multissite ativo-ativo atende ao tráfego de todas as regiões onde está implantado. Os clientes podem selecionar esta estratégia por outros motivos, além da DR. Ele pode ser usado para aumentar a disponibilidade ou ao implantar uma workload para um público global (para aproximar o endpoint dos usuários e/ou implantar pilhas localizadas para o público nessa região). Como uma estratégia de DR, se a workload não puder ser suportada em uma das Regiões da AWS onde está implantada, esta região será evacuada e as regiões restantes serão usadas para manter a disponibilidade. O multissite ativo-ativo é a estratégia de DR mais complexa operacionalmente e deve ser selecionada apenas quando os requisitos de negócios exigirem.

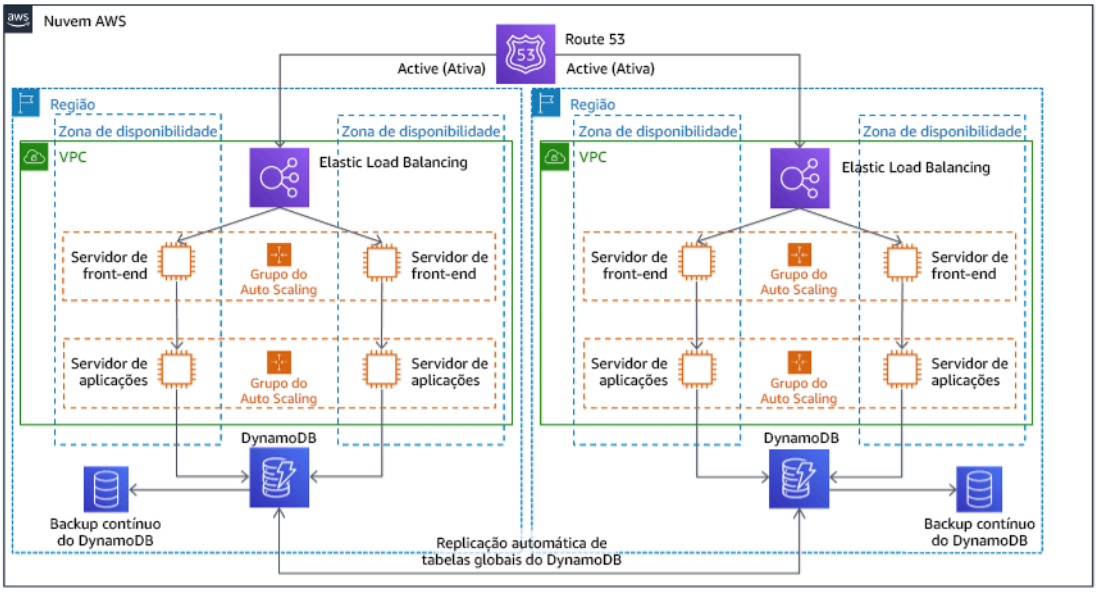


Figura 22: Arquitetura de multissite ativo-ativo

Para obter mais detalhes sobre esta estratégia, consulte Arquitetura de recuperação de desastres (DR) na AWS, parte IV: multissite ativo-ativo.

Práticas adicionais para proteção de dados

Com todas as estratégias, você também deve atenuar um desastre de dados. A replicação contínua de dados protege você contra alguns tipos de desastre, mas não contra corrupção ou destruição de dados, a menos que sua solução também inclua o versionamento de dados armazenados ou opções para recuperação a um ponto anterior no tempo. Você também deve fazer backup dos dados replicados no local de recuperação para criar backups pontuais além das réplicas.

O uso de várias zonas de disponibilidade (AZs) em uma única Região da AWS

Ao utilizar várias AZs em uma única região, sua implementação de DR usa vários elementos das estratégias acima. Primeiro, você deve criar uma arquitetura de alta disponibilidade (HA), usando várias AZs, conforme mostrado na figura 23. Esta arquitetura faz uso de uma abordagem multissite ativo-ativo, já que as instâncias do Amazon EC2 e a seção Elastic Load Balancer têm recursos implantados em várias AZs, processando solicitações ativamente. A arquitetura também demonstra standby a quente, no qual se a instância primária do Amazon RDS falhar (ou se a própria AZ falhar), a instância em espera será promovida a primária.

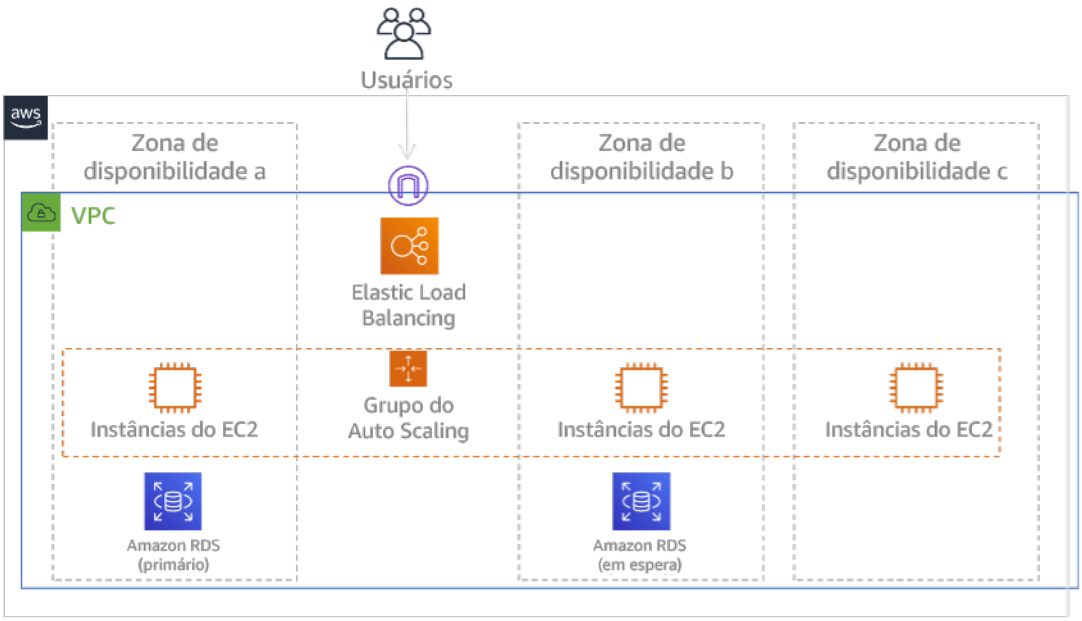


Figura 23: Arquitetura multi-A

Além da arquitetura de alta disponibilidade, você precisa adicionar backups de todos os dados necessários para executar a workload. Isso é especialmente importante para dados restritos a uma única zona, como volumes do Amazon EBS ou clusters do Amazon Redshift. Se uma AZ falhar, você precisará restaurar esses dados para outra AZ. Sempre que possível, você também deve copiar backups de dados para outra Região da AWS, como uma camada adicional de proteção.

Uma alternativa de abordagem menos comum para região única, DR multi-AZ é ilustrada na publicação do blog, Criação de aplicações altamente resilientes usando o Amazon Route 53 Application Recovery Controller, parte 1: pilha de região única. Aqui, a estratégia é manter o máximo de isolamento possível entre as AZs, assim como as regiões operam. Ao usar esta estratégia alternativa, você pode escolher uma abordagem ativa/ativa ou ativa/passiva.

Observação: algumas workloads têm requisitos regulamentares de residência de dados. Se isso se aplicar à sua workload em uma localidade que atualmente tem apenas uma Região da AWS, a multirregião não atenderá às suas necessidades de negócios. As estratégias multi-AZ fornecem boa proteção contra a maioria dos desastres.

3. Avalie os recursos da workload e qual será sua configuração na região de recuperação antes do failover (durante a operação normal).

Para infraestrutura e recursos da AWS, use a infraestrutura como código, como o AWS CloudFormation ou ferramentas de terceiros como o Hashicorp Terraform. Para implantar em várias

contas e regiões com uma única operação, você pode usar o AWS CloudFormation StackSets. Para estratégias multissite ativo-ativo e standby a quente, a infraestrutura implantada na região de recuperação tem os mesmos recursos que a região primária. Para as estratégias luz piloto e modo de espera passivo, a infraestrutura implantada exigirá ações adicionais para ficar pronta para produção. Ao usar o CloudFormation parâmetros e a lógica condicional, você pode controlar se uma pilha implantada está ativa ou em espera com um único modelo. Um exemplo desse modelo do CloudFormation está incluído nesta publicação do blog.

Todas as estratégias de DR exigem que sejam feitos backup das fontes de dados dentro da Região da AWS e, em seguida, esses backups sejam copiados para a região de recuperação. AWS Backup fornece uma visão centralizada na qual você pode configurar, programar e monitorar backups para esses recursos. Para luz piloto, modo de espera passivo e multissite ativo-ativo, você também deve replicar dados da região primária para recursos de dados na região de recuperação, como instâncias de banco de dados do Amazon Relational Database Service (Amazon RDS) ou tabelas do Amazon DynamoDB . Esses recursos de dados estão ativos e prontos para atender a solicitações na região de recuperação.

Para saber mais sobre como os serviços da AWS operam entre as regiões, consulte esta série de blogs em Criação de uma aplicação multirregional com os serviços da AWS.

4. Determine e implemente como deixar sua região de recuperação pronta para failover quando necessário (durante um evento de desastre).

Para multissite ativo-ativo, failover significa evacuar uma região e confiar nas regiões ativas restantes. No geral, essas regiões estão prontas para aceitar tráfego. Para as estratégias luz piloto e modo de espera passivo, as ações de recuperação precisarão implantar os recursos ausentes, como as instâncias do EC2 na figura 20, além de quaisquer outros recursos ausentes.

Para todas as estratégias acima, pode ser necessário promover instâncias somente leitura de bancos de dados para se tornar a instância primária de leitura/gravação.

Para backup e restauração, a restauração de dados do backup cria recursos para esses dados, como volumes do EBS, instâncias de banco de dados do RDS e tabelas do DynamoDB. Você também precisa restaurar a infraestrutura e implantar o código. É possível usar o AWS Backup para restaurar dados na região de recuperação. Perceber REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes para obter mais detalhes. A reconstrução da infraestrutura inclui a criação de recursos como instâncias do EC2, além do Amazon Virtual Private Cloud (Amazon VPC), sub-redes e grupos de segurança necessários. Você pode

automatizar grande parte do processo de restauração. Para saber mais, consulte nesta publicação do blog.

5. Determine e implemente como redirecionar o tráfego para failover quando necessário (durante um evento de desastre).

Essa operação de failover pode ser iniciada automaticamente ou manualmente. O failover iniciado automaticamente com base em verificações de integridade ou alarmes deve ser usado com cautela, pois um failover desnecessário (alarme falso) resulta em custos como indisponibilidade e perda de dados. Portanto, o failover iniciado manualmente é geralmente usado. Nesse caso, você ainda deve automatizar as etapas para failover, para que a inicialização manual seja como apertar um botão.

Há várias opções de gerenciamento de tráfego a serem consideradas ao usar os serviços da AWS. Uma opção é usar o Amazon Route 53. Ao usar o Amazon Route 53, você pode associar vários endpoints de IP em uma ou mais Regiões da AWS a um nome de domínio do Route 53. Para implementar o failover iniciado manualmente, você pode usar o Amazon Route 53 Application Recovery Controller, que fornece uma API de plano de dados altamente disponível para redirecionar o tráfego para a região de recuperação. Ao implementar o failover, use as operações do plano de dados e evite as do ambiente de gerenciamento, conforme descrito em REL11-BP04 Confiar no plano de dados e não no ambiente de gerenciamento durante a recuperação.

Para saber mais sobre esta e outras opções, consulte a seção de whitepaper sobre a recuperação de desastres.

6. Projete um plano de como a workload retornará.

Failback é quando você retorna a operação de workload para a região primária, após a redução de um evento de desastre. O provisionamento de infraestrutura e código para a região primária geralmente segue as mesmas etapas que foram usadas inicialmente, contando com a infraestrutura como código e pipelines de implantação de código. O desafio com o failback é restaurar os armazenamentos de dados e garantir sua consistência com a região de recuperação em operação.

No estado de failover, os bancos de dados na região de recuperação estão ativos e têm dados atualizados. O objetivo é ressincronizar da região de recuperação para a região primária, garantindo que ela esteja atualizada.

Alguns serviços da AWS farão isso automaticamente. Se usar as tabelas globais do Amazon DynamoDB, mesmo que a tabela na região primária tenha ficado indisponível, quando ela

voltar a ficar online, o DynamoDB retomará a propagação das gravações pendentes. Se usar o banco de dados global do Amazon Aurora e o failover planejado e gerenciado, a topologia de replicação existente do banco de dados global do Aurora é mantida. Portanto, a antiga instância de leitura/gravação na região primária se tornará uma réplica e receberá atualizações da região de recuperação.

Em casos onde isso não for automático, você precisará restabelecer o banco de dados na região primária como uma réplica do banco de dados na região de recuperação. Em muitos casos, isso envolverá a exclusão do banco de dados primário antigo e a criação de novas réplicas. Por exemplo, para obter instruções sobre como fazer isso com o banco de dados global do Amazon Aurora presumindo um failover não planejado, consulte este laboratório: Retornar um banco de dados global.

Após um failover, se você puder continuar executando na região de recuperação, considere torná-la a nova região primária. Você ainda seguiria todas as etapas acima para transformar a antiga região primária em uma região de recuperação. Algumas organizações fazem uma rotação programada, trocando suas regiões primárias e de recuperação periodicamente (por exemplo, a cada três meses).

Todas as etapas necessárias para failover e failback devem ser mantidas em um playbook disponível para todos os membros da equipe e que seja revisado periodicamente.

Nível de esforço para o plano de implementação: alto

Recursos

Práticas recomendadas relacionadas:

- the section called “REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes”
- the section called “REL11-BP04 Confiar no plano de dados e não no ambiente de gerenciamento durante a recuperação”
- the section called “REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados”

Documentos relacionados:

- Blog de arquitetura da AWS: série de recuperação de desastres
- Recuperação de desastres de workloads na AWS: recuperação na nuvem (whitepaper da AWS)

- Opções de recuperação de desastres na nuvem
- Crie uma solução de backend ativo-ativo multirregional sem servidor em uma hora
- Backend multirregional sem servidor: recarregado
- RDS: replicação de uma réplica de leitura entre regiões
- Route 53: configuração do failover de DNS
- S3: replicação entre regiões
- O que é o AWS Backup?
- O que é o Route 53 Application Recovery Controller?
- AWS Elastic Disaster Recovery
- HashiCorp Terraform: Introdução; AWS
- Parceiro do APN: parceiros que podem ajudar com a recuperação de desastres
- AWS Marketplace: produtos que podem ser usados para recuperação de desastres

Vídeos relacionados:

- Recuperação de desastres de workloads na AWS
- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
- Introdução ao AWS Elastic Disaster Recovery | Amazon Web Services

Exemplos relacionados:

- Laboratórios do AWS Well-Architected: recuperação de desastres : série de workshops que ilustram as estratégias de DR

REL13-BP03 Testar a implementação de recuperação de desastres para validá-la

Teste regularmente o failover no site de recuperação para garantir a operação adequada e que o RTO e o RPO sejam atendidos.

Um padrão que deve ser evitado é o desenvolvimento de caminhos de recuperação que raramente são executados. Por exemplo, você pode ter um repositório de dados secundário utilizado para consultas somente leitura. Quando você grava em um repositório de dados e o repositório de

dados primário falha, pode ser necessário fazer o failover para o repositório de dados secundário. Se você não testar esse failover com frequência, poderá descobrir que suas suposições sobre as capacidades do armazenamento de dados secundário são incorretas. A capacidade do secundário, que talvez tenha sido suficiente quando testado pela última vez, pode não conseguir mais tolerar a carga neste cenário. Nossa experiência mostrou que a única recuperação de erro que funciona é o caminho que você testa com frequência. É por isso que é melhor ter um pequeno número de caminhos de recuperação. Você pode estabelecer padrões de recuperação e testá-los regularmente. Se você tiver um caminho de recuperação complexo ou crítico, ainda precisará executar regularmente essa falha na produção para garantir o funcionamento desse caminho. No exemplo que acabamos de discutir, você deve realizar o failover para o standby regularmente, não importa a necessidade.

Antipadrões comuns:

- Nunca execute failovers em produção.

Benefícios do estabelecimento desta prática recomendada: Teste regularmente seu plano de recuperação de desastres para garantir que ele funcione quando necessário e que sua equipe saiba como executar a estratégia.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

- Projete suas cargas de trabalho para recuperação. Teste regularmente os caminhos de recuperação. A computação orientada à recuperação identifica as características nos sistemas que aprimoram a recuperação. Essas características são: isolamento e redundância, capacidade de reverter alterações em todo o sistema, capacidade de monitorar e determinar a integridade, capacidade de realizar diagnósticos, recuperação automatizada, design modular e recurso de reinicialização. Pratique o caminho de recuperação para garantir que possa realizá-la no tempo especificado para o estado determinado. Use seus runbooks durante essa recuperação para documentar problemas e encontrar soluções para eles antes do próximo teste.
  - O projeto de computação orientado por recuperação de Berkeley/Stanford
- Use o CloudEndure Disaster Recovery para implementar e testar sua estratégia de DR.
  - Teste da solução de recuperação de desastres com o CloudEndure
  - CloudEndure Disaster Recovery
  - CloudEndure Disaster Recovery para AWS

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar com a recuperação de desastres
- Blog de arquitetura da AWS: série de recuperação de desastres
- AWS Marketplace: produtos que podem ser usados para recuperação de desastres
- CloudEndure Disaster Recovery
- Recuperação de desastres de workloads na AWS: recuperação na nuvem (whitepaper da AWS)
- Teste da solução de recuperação de desastres com o CloudEndure
- O projeto de computação orientado por recuperação de Berkeley/Stanford
- O que é o AWS Fault Injection Simulator?

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
- AWS re:Invent 2019: Backup-and-restore and disaster-recovery solutions with AWS (STG208)

Exemplos relacionados:

- Laboratórios do AWS Well-Architected: testes de resiliência

REL13-BP04 Gerenciar 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.

O AWS Config monitora e registra continuamente as configurações dos recursos da AWS. Ele pode detectar desvios e acionar o AWS Systems Manager Automation para corrigi-lo e gerar alarmes. O AWS CloudFormation também pode detectar desvios nas pilhas que você implantou.

Antipadrões comuns:

- Falhar ao atualizar os locais de recuperação, ao fazer alterações de configuração ou infraestrutura nos locais primários.

- Não considerar possíveis limitações (como diferenças de serviço) nos locais primários e de recuperação.

Benefícios do estabelecimento desta prática recomendada: Garantir que o ambiente de DR seja consistente com seu ambiente existente para assegurar a recuperação completa.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Garanta que seus pipelines de entrega enviem para seus locais primário e de backup. Os pipelines de entrega para implantação de aplicativos em produção devem ser distribuídos para todos os locais de estratégia de recuperação de desastres especificados, incluindo os ambientes de desenvolvimento e de teste.
- Habilite o AWS Config para acompanhar possíveis locais de desvio. Use as regras do AWS Config para criar sistemas que aplicam suas estratégias de recuperação de desastres e geram alertas ao detectar desvios.
  - Correção de recursos não compatíveis do Regras do AWS Config pela AWS
  - AWS Systems Manager Automation
- Use o AWS CloudFormation para implantar a infraestrutura. O AWS CloudFormation pode detectar desvios entre as especificações dos modelos do CloudFormation e o que é realmente implantado.
  - AWS CloudFormation: detectar desvios em uma pilha inteira do CloudFormation

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar com a recuperação de desastres
- Blog de arquitetura da AWS: série de recuperação de desastres
- AWS CloudFormation: detectar desvios em uma pilha inteira do CloudFormation
- AWS Marketplace: produtos que podem ser usados para recuperação de desastres
- AWS Systems Manager Automation
- Recuperação de desastres de workloads na AWS: recuperação na nuvem (whitepaper da AWS)
- Como faço para implementar uma solução de gerenciamento de configuração de infraestrutura na AWS?

- Correção de recursos não compatíveis do Regras do AWS Config pela AWS

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)

REL13-BP05 Automatizar 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.

Com base em verificações de integridade configuradas, os serviços da AWS, como o Elastic Load Balancing e o AWS Auto Scaling, podem distribuir a carga para zonas de disponibilidade íntegras, enquanto outros serviços, como o Amazon Route 53 e o AWS Global Accelerator, podem rotear a carga para Regiões da AWS íntegras. O Amazon Route 53 Application Recovery Controller ajuda a gerenciar e coordenar o failover usando verificações de prontidão e recursos de controle de roteamento. Esses recursos monitoram continuamente a capacidade da aplicação de se recuperar de falhas, permitindo que você controle a recuperação da aplicação em várias Regiões da AWS, zonas de disponibilidade e ambientes on-premises.

Para workloads em datacenters físicos ou virtuais existentes ou nuvens privadas, o AWS Elastic Disaster Recovery, disponível por meio do AWS Marketplace, permite que as organizações configurem uma estratégia automatizada de recuperação de desastres para a AWS. O CloudEndure também oferece suporte à recuperação de desastres entre regiões e entre AZs na AWS.

Antipadrões comuns:

- A implementação de failover e failback automatizados idênticos pode causar oscilação quando uma falha ocorre.

Benefícios do estabelecimento dessa prática recomendada: A recuperação automatizada reduz o tempo de recuperação ao eliminar a oportunidade de erros manuais.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Automatize caminhos de recuperação. No caso de tempos de recuperação curtos, não é possível adotar critério e ação humanos em cenários de alta disponibilidade. O sistema deve recuperar-se automaticamente sob qualquer situação.
  - Use o CloudEndure Disaster Recovery para automatizar failover e failback. Ele replica continuamente suas máquinas (incluindo sistema operacional, configuração de estado do sistema, bancos de dados, aplicações e arquivos) em uma área de preparação de baixo custo na Conta da AWS de destino e na região de preferência. Em caso de desastre, você pode instruir o CloudEndure Disaster Recovery a executar automaticamente milhares de máquinas em seu estado totalmente provisionado em minutos.
    - Realizar um failover e failback de recuperação de desastres
    - CloudEndure Disaster Recovery

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar com a recuperação de desastres
- Blog de arquitetura da AWS: série de recuperação de desastres
- AWS Marketplace: produtos que podem ser usados para recuperação de desastres
- AWS Systems Manager Automation
- CloudEndure Disaster Recovery para AWS
- Recuperação de desastres de workloads na AWS: recuperação na nuvem (whitepaper da AWS)

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)

Eficiência de performance

Tópicos

- Seleção
- Análise

- Monitoramento
- Concessões

Seleção

Perguntas

- PERF 1 Como você seleciona a arquitetura de melhor performance?
- PERF 2 Como você seleciona sua solução de computação?
- PERF 3 Como você seleciona sua solução de armazenamento?
- PERF 4 Como você seleciona sua solução de banco de dados?
- PERF 5 Como você configura sua solução de rede?

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.

Práticas recomendadas

- PERF01-BP01 Compreender os serviços e os recursos disponíveis
- PERF01-BP02 Definir um processo para as opções de arquitetura
- PERF01-BP03 Fatorar os requisitos de custo ao tomar decisões
- PERF01-BP04 Usar políticas ou arquiteturas de referência
- PERF01-BP05 Usar as orientações do seu provedor de nuvem ou de um parceiro apropriado
- PERF01-BP06 Realizar testes comparativos das workloads
- PERF01-BP07 Teste de carga da workload

PERF01-BP01 Compreender 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.

Caso esteja avaliando uma carga de trabalho existente, é necessário gerar um inventário dos vários recursos de serviços que ela consome. Seu inventário ajuda a avaliar quais componentes podem ser substituídos por serviços gerenciados e tecnologias mais recentes.

Antipadrões comuns:

- Você usa a nuvem como um datacenter colocalizado.
- Você usa o armazenamento compartilhado para todas as coisas que precisam de armazenamento persistente.
- Você não usa a escalabilidade automática.
- Você usa tipos de instância mais próximos aos padrões atuais, mas maiores, quando necessário.
- Você implanta e gerencia tecnologias disponíveis como serviços gerenciados.

Benefícios do estabelecimento desta prática recomendada: Ao considerar os serviços com os quais você não está familiarizado, você pode reduzir significativamente o custo da infraestrutura e o esforço necessário para manter seus serviços. Você pode acelerar seu tempo de entrada no mercado implantando novos serviços e recursos.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

Fazer o inventário do software da workload e da arquitetura dos serviços relacionados: colete um inventário da sua workload e decida sobre qual categoria de produtos saber mais. Identifique os componentes da workload que podem ser substituídos por serviços gerenciados para melhorar a performance e reduzir a complexidade operacional.

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- This is my Architecture

Exemplos relacionados:

- AWS Samples (Amostras da AWS)
- AWS SDK Examples (Exemplos do AWS SDK)

PERF01-BP02 Definir um processo para as opções de arquitetura

Use a experiência e o conhecimento internos da nuvem ou os 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.

Ao escrever histórias críticas de usuários para sua arquitetura, inclua os requisitos de performance, como especificar a rapidez em que cada história crítica deve ser executada. Para essas histórias essenciais, implemente jornadas de usuário em roteiros adicionais a fim de garantir a visibilidade da performance delas em relação aos seus requisitos.

Antipadrões comuns:

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

Benefícios do estabelecimento desta prática recomendada: Ao ter um processo definido para fazer alterações de arquitetura, você habilita o uso dos dados coletados para influenciar o design da carga de trabalho ao longo do tempo.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

Selecionar uma abordagem da arquitetura: identifique o tipo de arquitetura que atende aos seus requisitos de performance. Identifique restrições como o meio de entrega (desktop, web, dispositivo móvel, IoT), requisitos legados e integrações. Identifique oportunidades para reutilização, incluindo refatoração. Consulte outras equipes, diagramas de arquitetura e recursos, como os arquitetos de

solução da AWS, as arquiteturas de referência da AWS e os parceiros da AWS, para ajudá-lo a escolher uma arquitetura.

Definir métricas de performance: use a experiência do cliente para identificar as métricas mais importantes. Para cada métrica, identifique o alvo, a abordagem de medição e a prioridade. Defina a experiência do cliente. Documente a experiência de performance exigida pelos clientes, incluindo como os clientes julgarão a performance da carga de trabalho. Priorize questões de experiência para histórias de usuário importantes. Inclua requisitos de performance e implemente jornadas de usuários com script para garantir que você saiba como as histórias se comportam de acordo com seus requisitos.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de soluções da AWS
- Centro de Conhecimentos da AWS

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- This is my Architecture

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS

PERF01-BP03 Fatorar 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.

Determine quais componentes da workload podem ser substituídos por serviços totalmente gerenciados, como bancos de dados gerenciados, caches na memória e outros serviços de ETL.

A redução de sua carga de trabalho operacional permite que você concentre os recursos em resultados empresariais.

Para conhecer as melhores práticas de requisitos de custo, consulte a seção Recursos econômicos do whitepaper sobre o pilar de otimização de custos.

Antipadrões comuns:

- Você só usa uma família de instâncias.
- Você não avalia soluções licenciadas em relação a soluções de código aberto
- Você só usa o armazenamento em bloco.
- Implante software comum em instâncias do EC2 e volumes do Amazon EBS ou efêmeros disponíveis como um serviço gerenciado.

Benefícios do estabelecimento desta prática recomendada: Considerar o custo ao fazer suas escolhas permitirá que você habilite outros investimentos.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

Otimizar os componentes da workload para reduzir custos: dimensione os componentes da workload e habilite a elasticidade para reduzir custos e maximizar a eficiência dos componentes. Determine quais componentes da carga de trabalho podem ser substituídos por serviços gerenciados quando apropriado, como bancos de dados gerenciados, caches na memória e proxies reversos.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de soluções da AWS
- Centro de Conhecimentos da AWS
- AWS Compute Optimizer

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- This is my Architecture
- Optimize performance and cost for your AWS compute (CMP323-R1) (Otimizar a performance e os custos de sua computação da AWS (CMP323-R1))

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS
- Rightsizing with Compute Optimizer and Memory utilization enabled (Dimensionamento correto com o AWS Compute Optimizer e utilização da memória ativada)
- Código para demonstração do AWS Compute Optimizer

PERF01-BP04 Usar 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.

Antipadrões comuns:

- Você permite um amplo uso da escolha de tecnologia, que pode afetar a sobrecarga de gerenciamento da sua empresa.

Benefícios do estabelecimento desta prática recomendada: Estabelecer uma política para opções de arquitetura, tecnologia e fornecedor permitirá que as decisões sejam tomadas rapidamente.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

Implantar a workload usando políticas existentes ou arquiteturas de referência: integre os serviços à sua implantação de nuvem e use os testes de performance para garantir que seja possível continuar a atender aos requisitos de performance.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de soluções da AWS
- Centro de Conhecimentos da AWS

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- This is my Architecture

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS

PERF01-BP05 Usar 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.

Entre em contato com a AWS para obter assistência quando precisar de orientações ou informações adicionais sobre produtos. Os arquitetos de soluções da AWS e o AWS Professional Services fornecem orientação para a implementação da solução. Os parceiros da AWS oferecem toda a experiência na AWS para ajudar você a adquirir agilidade e inovação para os seus negócios.

Antipadrões comuns:

- Você usa a AWS como um provedor de datacenter comum.
- Você usa os serviços da AWS de uma maneira para a qual eles não foram projetados.

Benefícios do estabelecimento desta prática recomendada: Consultar o seu provedor ou um parceiro trará confiança para sua tomada de decisões.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

Entrar em contato com os recursos da AWS para obter assistência: os arquitetos de soluções e o AWS Professional Services fornecem orientações para a implementação de soluções. Os parceiros da APN fornecem a experiência na AWS para ajudar você a adquirir agilidade e inovação para a sua empresa.

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- This is my Architecture

Exemplos relacionados:

- AWS Samples (Amostras da AWS)
- AWS SDK Examples (Exemplos do AWS SDK)

PERF01-BP06 Realizar testes comparativos das workloads

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.

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

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

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

Como normalmente é mais rápido executar testes comparativos do que testes de carga, eles podem ser usados mais cedo no pipeline de implantação e fornecer um feedback mais rápido sobre desvios de performance. Ao avaliar uma alteração significativa em um componente ou serviço, um benchmark pode ser uma maneira rápida de verificar se é possível justificar a iniciativa para concretizar a alteração. O uso de benchmarking em conjunto com testes de carga é importante porque o teste de carga informa como será a performance de sua carga de trabalho em produção.

Antipadrões comuns:

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

Benefícios do estabelecimento desta prática recomendada: O benchmarking da sua implementação atual permite medir a melhoria da performance.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

Monitorar a performance durante o desenvolvimento: implemente processos que deem visibilidade da performance à medida que a workload evolui.

Integrar-se com o pipeline de entrega: execute testes de carga automáticos em seu pipeline de entrega. Compare os resultados do teste com indicadores-chave de performance (KPIs) e limites predefinidos para garantir que você continue atendendo aos requisitos de performance.

Testar a jornada dos usuários: use versões sintéticas ou limpas de dados de produção (remova informações confidenciais ou de identificação) para o teste de carga. Exercite toda sua arquitetura usando jornadas do usuário reproduzidas ou pré-programadas por meio de seu aplicativo em escala.

Monitoramento de usuários reais: use o CloudWatch RUM para ajudar a coletar e visualizar dados do lado do cliente sobre a performance da sua aplicação. Use esses dados para ajudar a estabelecer testes comparativos de performance de usuários reais.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de soluções da AWS
- Centro de Conhecimentos da AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- This is my Architecture
- Optimize applications through Amazon CloudWatch RUM (Otimizar as aplicações por meio do Amazon CloudWatch RUM)
- Demo of Amazon CloudWatch Synthetics (Demonstração do Amazon CloudWatch Synthetics)

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS
- Testes de carga distribuída
- Medição do tempo de carga da página com o Amazon CloudWatch Synthetics
- Cliente da web do Amazon CloudWatch RUM

PERF01-BP07 Teste de carga da workload

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.

Testes de carga usam sua carga de trabalho real para que você possa verificar a performance da solução em um ambiente de produção. Execute os testes de carga usando versões sintéticas ou limpas dos dados de produção (remova informações confidenciais ou de identificação). Empregue reproduções ou pré-programações de jornadas de usuário em escala em sua carga de trabalho para exercitar toda a sua arquitetura. Realize testes de carga automaticamente como parte de seu pipeline de entrega e compare os resultados a Key Performance Indicators (KPI – Indicadores-chave de performance) e limites predefinidos. Isso garante que você continue a alcançar a performance necessária.

Antipadrões comuns:

- Você realiza um teste de carga de peças individuais da carga de trabalho, mas não toda a carga de trabalho.
- Você realiza um teste de carga em uma infraestrutura que não é igual ao seu ambiente de produção.
- Você só realiza testes de carga para a carga esperada e não para além dela, para ajudar a prever onde você pode ter problemas futuros.
- Executar testes de carga sem informar o AWS Support e ter o teste anulado por parecer um evento de negação de serviço.

Benefícios do estabelecimento desta prática recomendada: Medir sua performance em um teste de carga mostrará onde você será afetado à medida que a carga aumentar. Com isso você terá a capacidade de antecipar as alterações necessárias antes que elas afetem sua carga de trabalho.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

Validar sua abordagem com testes de carga: o teste de carga é uma comprovação de conceito para descobrir se você atende aos requisitos de performance. É possível usar os serviços da AWS para executar ambientes em escala de produção para testar a arquitetura. Como você paga apenas pelo ambiente de teste quando precisa usá-lo, é possível realizar um teste em escala total a um custo bem menor do que usando um ambiente no local.

Monitorar métricas: o Amazon CloudWatch pode coletar métricas entre os recursos da sua arquitetura. Você também pode coletar e publicar métricas personalizadas para descobrir métricas

de negócio ou derivadas. Use o CloudWatch ou soluções de terceiros para definir alarmes que indiquem quando os limites são violados.

Testar em escala: os testes de carga usam a workload real para que você possa verificar a performance da solução em um ambiente de produção. É possível usar os serviços da AWS para executar ambientes em escala de produção para testar a arquitetura. Como você apenas paga pelo ambiente de teste quando ele é necessário, pode realizar um teste em escala total a um custo menor do que usando um ambiente no local. Aproveite o Nuvem AWS para testar a workload para descobrir se há uma falha na escala ou se ela está com a escala reduzida horizontalmente de maneira não linear. Por exemplo, use instâncias spot para gerar cargas a um baixo custo e descobrir gargalos antes que eles ocorram em produção.

Recursos

Documentos relacionados:

- AWS CloudFormation
- Building AWS CloudFormation Templates using CloudFormer (Criação de modelos do AWS CloudFormation usando o CloudFormer)
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Testes de carga distribuída na AWS

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- Optimize applications through Amazon CloudWatch RUM (Otimizar as aplicações por meio do Amazon CloudWatch RUM)
- Demo of Amazon CloudWatch Synthetics (Demonstração do Amazon CloudWatch Synthetics)

Exemplos relacionados:

- Testes de carga distribuída na AWS

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.

Práticas recomendadas

- PERF02-BP01 Avaliar as opções de computação disponíveis
- PERF02-BP02 Compreender as opções de configuração de computação disponíveis
- PERF02-BP03 Coletar métricas relacionadas à computação
- PERF02-BP04 Determinar a configuração necessária com o dimensionamento correto
- PERF02-BP05 Usar a elasticidade dos recursos disponíveis
- PERF02-BP06 Reavaliar as necessidades de computação com base em métricas

PERF02-BP01 Avaliar as opções de computação disponíveis

Compreenda como a workload pode se beneficiar do uso de diferentes opções de computação, como instâncias, contêineres e funções.

Resultado desejado: Ao compreender todas as opções de computação disponíveis, você conhecerá oportunidades para aumentar a performance, reduzir custos desnecessários de infraestrutura e diminuir o esforço necessário para manter a workload. Você também pode acelerar seu tempo de entrada no mercado ao implantar novos serviços e recursos.

Antipadrões comuns:

- Em uma workload de pós-migração, usar a mesma solução de computação que foi usada on-premises.
- Falta de conhecimento das soluções de computação de nuvem e como essas soluções podem melhorar a performance da computação.
- Superdimensionar uma solução de computação existente para atender aos requisitos de escalabilidade ou de performance, quando uma solução de computação alternativa se alinharia com as características da sua workload de forma mais exata.

Benefícios do estabelecimento desta prática recomendada: Ao identificar os requisitos de computação e avaliar as soluções de computação disponíveis, as partes interessadas e as equipes de engenharia da empresa compreenderão os benefícios e as limitações de usar a solução de computação selecionada. A solução de computação selecionada deve se ajustar aos critérios de performance da workload. Os principais critérios incluem as necessidades de processamento, os padrões do tráfego, os padrões de acesso aos dados e as necessidades de escalabilidade e de latência.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

Compreenda as soluções de virtualização, conteinerização e gerenciamento que podem beneficiar sua workload e atender aos requisitos de performance. Uma workload pode conter vários tipos de soluções de computação. Cada solução de computação tem características diferentes. Com base na escala da sua workload e nos requisitos de computação, uma solução de computação pode ser selecionada e configurada para atender às suas necessidades. O arquiteto de nuvem deve aprender as vantagens e as desvantagens de instâncias, contêineres e funções. As seguintes etapas ajudarão você a selecionar a solução de computação que corresponda às características de sua workload e requisitos de performance.

Tipo	Servidor	Contêineres	Função
Serviço da AWS	Amazon Elastic Compute Cloud (Amazon EC2)	Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	AWS Lambda
Características principais	Tem opção dedicada para requisitos de licenças de hardware, opções de posicionamento e uma grande seleção de diferentes famílias de instâncias com	Implantação fácil, ambientes consistentes, execuções em instâncias do EC2, escalável	Tempo limite curto (15 minutos ou menos), memória e CPU máximas não são tão altas quanto para outros serviços, camada de hardware gerenciada, escala

Tipo	Servidor	Contêineres	Função
	base em métricas de computação		para milhões de solicitação simultâne as
Casos de uso comuns	Migrações do tipo mover sem alteraçõe s (lift-and-shift), aplicações monolític as, ambientes híbridos, aplicações empresariais	Microsserviços, ambientes híbridos,	Microsserviços, aplicações orientadas por eventos

Etapas da implementação:

1. Para selecionar o local em que a solução de computação deve residir, avalie a the section called "PERF05-BP06 Escolher o local da sua workload com base nos requisitos de rede". O local limitará os tipos de solução de computação disponíveis para você.
2. Identifique o tipo de solução de computação que funciona com os requisitos de local e das aplicações
   a. Amazon Elastic Compute Cloud (Amazon EC2) instâncias de servidor virtual são fornecidas em uma grande variedade de famílias e de tamanhos diferentes. Elas oferecem uma grande variedade de capacidades, como unidades de estado sólido (SSDs) e unidades de processamento gráfico (GPUs). As instâncias do EC2 oferecem a maior flexibilidade de opções de instâncias. Ao iniciar uma instância do EC2, o tipo de instância especificado determina o hardware da instância. Cada tipo de instância oferece diferentes capacidades de computação, memória e armazenamento. Os tipos de instância são agrupados em famílias de instância conforme essas capacidades. Os casos de uso típicos incluem: execução de aplicações empresariais, computação de alta performance (HPC), treinamento e implantação de aplicações de machine learning e execução de aplicações nativas de nuvem.
   b. Amazon Elastic Container Service (Amazon ECS) é um serviço totalmente gerenciado de orquestração de contêineres que permite executar e gerenciar automaticamente contêineres em um cluster de instâncias do EC2 ou instâncias de tecnologia sem servidor usando o AWS Fargate. É possível usar o Amazon ECS com outros serviços, como o Amazon Route 53, o Secrets Manager, o AWS Identity and Access Management (IAM), e o Amazon CloudWatch. O

Amazon ECS é recomendado quando sua aplicação é conteinerizada e a equipe de engenharia prefere contêineres do Docker.

c. Amazon Elastic Kubernetes Service (Amazon EKS) é um serviço gerenciado do Kubernetes. É possível optar por executar os clusters do EKS usando o AWS Fargate, eliminando a necessidade de provisionar e de gerenciar servidores. O gerenciamento do Amazon EKS é simplificado devido às integrações com os serviços da AWS, como o Amazon CloudWatch, os grupos do Auto Scaling, o AWS Identity and Access Management (IAM) e a Amazon Virtual Private Cloud (VPC). Ao usar contêineres, use métricas de computação para selecionar o tipo ideal para a sua workload, de forma semelhante a como você usa métricas de computação para selecionar seus tipos de instância do EC2 ou do AWS Fargate. O Amazon EKS é recomendado quando sua aplicação é conteinerizada e a equipe de engenharia prefere contêineres do Kubernetes em vez dos contêineres do Docker.

d. Você pode usar o AWS Lambda para executar código compatível com as opções permitidas de tempo de execução, memória e CPU. Basta fazer upload do seu código, e o AWS Lambda gerenciará tudo o que for necessário para executar e ajustar a escala desse código. É possível configurar o código para ser acionado automaticamente em outros serviços da AWS ou chamado diretamente. O Lambda é recomendado para execuções curtas, arquiteturas de microsserviço desenvolvidas para a nuvem.

3. Depois de experimentar a nova solução de computação, planeje a migração e valide as métricas de performance. Esse é um processo contínuo, consulte a the section called “PERF02-BP04 Determinar a configuração necessária com o dimensionamento correto”.

Nível de esforço para o plano de implementação: Se uma workload estiver sendo movida de uma solução de computação para outra, poderá haver um nível moderado de esforço envolvido na refatoração da aplicação.

Recursos

Documentos relacionados:

- Computação em nuvem com a AWS
- Tipos de instância do EC2
- Controle do estado do processo para sua instância do EC2
- Contêineres do EKS: nós de processamento do EKS
- Contêineres do Amazon ECS: instâncias de contêineres do Amazon ECS
- Funções: configuração de funções do Lambda

- Prescriptive Guidance for Containers (Orientações prescritivas para contêineres)
- Prescriptive Guidance for Serverless (Orientações prescritivas para a tecnologia sem servidor)

Vídeos relacionados:

- Como escolher uma opção de computação para startups
- Optimize performance and cost for your AWS compute (CMP323-R1) (Otimizar a performance e os custos de sua computação da AWS (CMP323-R1))
- Amazon EC2 foundations (CMP211-R2)
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Deliver high performance ML inference with AWS Inferentia (CMP324-R1) (Entregar inferência de ML de alta performance com o AWS Inferentia (CMP324-R1))
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (CMP202-R1) (Computação melhor, mais rápida e com custo mais baixo: otimização de custos com o EC2 (CMP202-R1))

Exemplos relacionados:

- Migração de aplicações da web para contêineres
- Executar uma aplicação Hello World de tecnologia sem servidor

PERF02-BP02 Compreender as opções de configuração de computação disponíveis

Cada solução de computação tem opções e configurações disponíveis para oferecer suporte às características da sua workload. Saiba como várias opções complementam a workload, e quais opções de configuração são melhores para a sua aplicação. Exemplos dessas opções são famílias de instâncias, tamanhos, recursos (GPU, E/S), expansão, tempos limite, tamanhos de função, instâncias de contêineres e simultaneidade.

Resultado desejado: As características da workload, incluindo CPU, memória, throughput da rede, GPU, IOPs, padrões de tráfego e padrões de acesso aos dados, são documentadas e usadas para configurar a solução de computação que corresponda a essas características. Cada uma dessas métricas, além das métricas personalizadas específicas da sua workload, são registradas, monitoradas e usadas para otimizar a configuração da computação que melhor atenda às suas necessidades.

Antipadrões comuns:

- Usar a mesma solução de computação que foi usada on-premises.
- Não analisar as opções de computação ou as famílias de instâncias que atendam às características da workload.
- Superdimensionar a computação para garantir a capacidade de expansão.
- Você usa várias plataformas de gerenciamento de computação para a mesma carga de trabalho.

Benefícios do estabelecimento desta prática recomendada: Familiarize-se com as ofertas de computação da AWS para determinar a solução correta para cada uma das suas workloads. Depois de selecionar as ofertas de computação para a workload, é possível experimentá-las rapidamente para determinar se elas atendem às necessidades da sua workload. Uma solução de computação otimizada para atender às características da sua workload melhorará a sua performance, reduzirá os seus custos e aumentará a sua confiabilidade.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

Se a sua workload estiver usando a mesma opção de computação há mais de quatro semanas, e se a previsão for de que as características permanecerão as mesmas no futuro, você poderá usar o AWS Compute Optimizer para obter uma recomendação com base nas características de sua computação. Se o AWS Compute Optimizer não for uma opção por causa da falta de métricas, de um tipo de instância não compatível ou de uma alteração previsível em suas características, preveja suas métricas com base nos testes e experimentação de carga.

Etapas da implementação:

1. Você está executando em instâncias ou contêineres do EC2 com o tipo de execução do EC2?
    a. Sua workload pode usar CPUs para melhorar a performance?
        i. Instâncias de computação acelerada são instâncias baseadas em GPU que fornecem a mais alta performance para treinamento de machine learning, inferência e computação de alta performance.
    b. Sua workload executa aplicações de inferência de machine learning?
        i. AWS Inferentia (Inf1) as instâncias do Inf1 são criadas para serem compatíveis com aplicações de inferência de machine learning. Ao usar instâncias do Inf1, os clientes podem executar aplicações de inferência de machine learning em grande escala, como reconhecimento de imagens, reconhecimento de fala, processamento de linguagem natural, personalização e detecção de fraude. É possível criar um modelo em umas das principais

frameworks de machine learning, como o TensorFlow, o PyTorch ou o MXNet, e usar instâncias de GPU para treinar o seu modelo. Depois que o modelo de machine learning for treinado para atender aos seus requisitos, você poderá implantá-lo em instâncias Inf1 usando o AWS Neuron,um kit de desenvolvimento de software (SDK) especializado que consiste em um compilador, um tempo de execução e de ferramentas de criação de perfil que otimizam a performance de inferência de machine learning de chips do Inferentia.

c. A sua workload se integra com hardware de baixo nível para melhorar a performance?

   i. Matrizes de porta programável no campo (FPGAs) Usando FPGAs, é possível otimizar as workloads com a execução acelerada por hardware personalizada para as workloads mais exigentes. É possível definir seus algoritmos usando linguagens gerais de programação compatíveis como C ou Go, ou linguagens orientadas por hardware, como Verilog ou VHDL.

d. Você tem pelo menos quatro semanas de métricas e pode prever se o padrão e as métricas do tráfego permanecerão iguais no futuro?

   i. Uso Compute Optimizer para obter uma recomendação de machine learning sobre a configuração de computação que corresponde melhor às características de sua computação.

e. A performance da sua workload é limitada pelas métricas de CPU?

   i. As instâncias otimizadas por computação são ideais para as workloads que exigem processadores de alta performance.

f. A performance de sua workload é limitada pelas métricas de memória?

   i. As instâncias otimizadas por memória entregam grandes quantidades de memória para oferecer suporte às workloads com consumo intenso de memória.

g. A performance de sua workload é limitada por IOPS?

   i. As instâncias otimizadas por armazenamento são projetadas para workloads que exigem alta leitura sequencial e acesso de gravação (IOPS) no armazenamento local.

h. As características da sua workload representam uma necessidade balanceada entre todas as métricas?

   i. A CPU da sua workload precisa de expansão para tratar picos no tráfego?

      A. As instâncias de performance expansível são semelhantes às instâncias otimizadas para computação, com a exceção de que oferecem a capacidade de expandir além da linha de base fixa da CPU identificada em uma instância otimizada para computação.

   ii. As instâncias de uso geral fornecem um equilíbrio de todas as características para compatibilidade com uma variedade de workloads.

   i. Sua instância de computação é executada no Linux e é restringida pelo throughput da rede na placa de interface da rede?
      i. Análise Pergunta sobre performance 5, Prática recomendada 2: avaliar os recursos de rede disponíveis para encontrar o tipo e a família de instâncias certos para atender às suas necessidades de performance.
   j. Sua workload precisa de instâncias consistentes e previsíveis em uma zona de disponibilidade específica que pode ser confirmada por um ano?
      i. Instâncias reservadas confirmam as reservas de capacidade em uma zona de disponibilidade específica. As instâncias reservadas são ideais para a capacidade de computação exigida em uma zona de disponibilidade específica.
   k. Sua workload tem licenças que exigem hardware dedicado?
      i. Hosts dedicados são compatíveis com licenças de software e ajudam você a atender aos requisitos de conformidade.
   l. Sua solução de computação se expande e exige processamento síncrono?
      i. Instâncias sob demanda permitem usar a capacidade de computação pela hora ou segundo sem uma confirmação de longo prazo. Essas instâncias são ideais para expansões acima das necessidades de performance da linha de base.
   m. Sua solução de computação é sem estado, tolerante à falhas e assíncrona?
      i. Instâncias spot permitem aproveitar a capacidade de instâncias não utilizadas para workloads sem estado e tolerantes à falhas.
2. Você executa contêineres no Fargate?
   a. A performance de suas tarefas é restringida pela memória ou pela CPU?
      i. Use a ferramenta de recomendações do Tamanho da tarefa para ajustar a memória ou a CPU.
   b. Sua performance está sendo afetada por expansões do seu padrão de tráfego?
      i. Use a ferramenta de recomendações do Auto Scaling configuração para corresponder seus padrões de tráfego.
3. Sua solução de computação é no Lambda?
   a. Você tem pelo menos quatro semanas de métricas e pode prever se o padrão e as métricas do tráfego permanecerão iguais no futuro?
      i. Uso Compute Optimizer para obter uma recomendação de machine learning sobre a configuração de computação que corresponde melhor às características de sua computação.
   b. Você não têm métricas suficientes para usar o AWS Compute Optimizer?

   i. Se você não tiver métricas disponíveis para usar o Compute Optimizer, use o Ajuste da potência do AWS Lambda para ajudar a selecionar a melhor configuração.

c. A performance da função é restringida pela memória ou pela CPU?
   i. Configure a memória do Lambda para atender às suas métricas de necessidades de performance.

d. O tempo limite de sua função está se esgotando ao executar?
   i. Altere as configurações de tempo limite

e. A performance de sua função é restringida pelas expansões de atividades e pela simultaneidade?
   i. Defina as configurações de simultaneidade para atender aos seus requisitos de performance.

f. Sua função é executada assincronamente e falha em novas tentativas?
   i. Configure a idade máxima do evento e o limite máximo de novas tentativas nas definições da configuração assíncrona.

Nível de esforço do plano de implementação:

Ao estabelecer esta prática recomendada, lembre-se das características e das métricas atuais da computação. A coleta dessas métricas, o estabelecimento de uma linha de base e o uso dessas métricas para identificar a opção ideal de computação é um nível de esforço baixo to moderado . Isso é melhor validade com testes de carga e experimentação.

Recursos

Documentos relacionados:

- Computação em nuvem com a AWS
- AWS Compute Optimizer
- Tipos de instância do EC2
- Controle do estado do processo para sua instância do EC2
- Contêineres do EKS: nós de processamento do EKS
- Contêineres do Amazon ECS: instâncias de contêineres do Amazon ECS
- Funções: configuração de funções do Lambda

Vídeos relacionados:

- Amazon EC2 foundations (CMP211-R2)
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Optimize performance and cost for your AWS compute (CMP323-R1) (Otimizar a performance e os custos de sua computação da AWS (CMP323-R1))

Exemplos relacionados:

- Rightsizing with Compute Optimizer and Memory utilization enabled (Dimensionamento correto com o AWS Compute Optimizer e utilização da memória ativada)
- Código para demonstração do AWS Compute Optimizer

PERF02-BP03 Coletar métricas relacionadas à computação

Para entender a performance dos recursos de computação, registre e acompanhe a utilização de vários sistemas. Esses dados podem ser usados para fazer determinações mais precisas sobre os requisitos de recursos.

As workloads podem gerar grandes volumes de dados, como métricas, logs e eventos. Determine se o serviço de armazenamento, monitoramento e observação existente é capaz de gerenciar os dados gerados. Identifique quais métricas refletem a utilização de recursos e podem ser coletadas, agregadas e correlacionadas em uma única plataforma. Essas métricas devem representar todos os recursos de workload, aplicações e serviços, para que você possa visualizar facilmente todo o sistema e identificar oportunidades e problemas na melhoria de performance.

Resultado desejado: todas as métricas referentes aos recursos relacionados à computação são identificadas, coletadas, agregadas e correlacionadas em uma única plataforma com retenção implementada para oferecer suporte a metas operacionais e de custo.

Antipadrões comuns:

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

Benefícios do estabelecimento dessa prática recomendada: para monitorar a performance das workloads, você precisa registrar várias métricas de performance ao longo de um período. Essas métricas permitem detectar anomalias na performance. Elas também ajudarão a avaliar a performance em relação às métricas de negócios para garantir que as necessidades da workload sejam atendidas.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

Identifique, colete, agregue e correlacione métricas relacionadas à computação. Usar um serviço como o Amazon CloudWatch pode tornar a implementação mais rápida e fácil de manter. Além das métricas padrão registradas, identifique e acompanhe métricas adicionais em nível de sistema na workload. Registre dados como níveis de uso da CPU, memória, E/S de disco e métricas de entrada e saída de rede para obter uma percepção sobre os níveis de utilização ou os gargalos. Esses dados são cruciais para entender a performance da workload e como a solução de computação é utilizada. Use essas métricas como parte de uma abordagem impulsionada por dados para ajustar e otimizar ativamente os recursos de sua carga de trabalho.

Etapas da implementação:

1. Quais métricas de solução de computação são importantes de acompanhar?
    a. Métricas padrão do EC2
    b. Métricas padrão do Amazon ECS
    c. Métricas padrão do EKS
    d. Métricas padrão do Lambda
    e. Métricas de memória e de disco do EC2
2. Tenho, atualmente, uma solução de registro em log e monitoramento aprovada?
    a. Amazon CloudWatch
    b. AWS Distro for OpenTelemetry
    c. Amazon Managed Service for Prometheus
3. Identifiquei e configurei minhas políticas de retenção de dados para corresponder às minhas metas operacionais e de segurança?
    a. Retenção de dados padrão para métricas do CloudWatch
    b. Retenção de dados padrão para o CloudWatch Logs
4. Como você implanta agentes de agregação de métrica e log?

a. AWS Systems Manager Automation
b. OpenTelemetry Collector

Nível de esforço para o plano de implementação: Há um nível de esforço médio para identificar, rastrear, coletar, agregar e correlacionar métricas de todos os recursos de computação.

Recursos

Documentos relacionados:

- Documentação do Amazon CloudWatch
- Coletar métricas e logs das instâncias do Amazon EC2 e de servidores on-premises com o agente do CloudWatch
- Acessar o Amazon CloudWatch Logs para o AWS Lambda
- Uso do CloudWatch Logs com instâncias de contêiner
- Publicar métricas personalizadas
- AWS Answers: registro em log centralizado
- Serviços da AWS que publicam métricas do CloudWatch
- Monitoramento do Amazon EKS no AWS Fargate

Vídeos relacionados:

- Application Performance Management na AWS
- Build a monitoring plan

Exemplos relacionados:

- Nível 100: monitoramento com os painéis do CloudWatch
- Nível 100: monitoramento das instâncias do Windows do EC2 com os painéis do CloudWatch
- Nível 100: monitoramento de uma instância do Amazon Linux EC2 com os painéis do CloudWatch

PERF02-BP04 Determinar a configuração necessária com 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.

Antipadrões comuns:

- Você escolhe a maior instância disponível para todas as cargas de trabalho.
- Você padroniza todos os tipos de instâncias para um tipo a fim de facilitar o gerenciamento.

Benefícios do estabelecimento desta prática recomendada: A familiarização com as ofertas de computação da AWS permite determinar a solução correta para suas várias workloads. Depois de selecionar as várias ofertas de computação para a workload, você terá agilidade para experimentar rapidamente essas ofertas de computação e determinar quais atendem às necessidades da sua workload.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

Modificar a configuração da sua workload por meio de dimensionamento correto: para otimizar a performance e a eficiência geral, determine de quais recursos sua workload precisa. Escolha instâncias com otimização de memória para sistemas que exigem mais memória que a CPU ou instâncias com otimização de computação para componentes que realizam processamento de dados que não consome muita memória. O dimensionamento correto habilita sua carga de trabalho a operar da melhor maneira possível enquanto consome apenas os recursos necessários

Recursos

Documentos relacionados:

- AWS Compute Optimizer
- Computação em nuvem com a AWS
- Tipos de instância do EC2

- Contêineres do ECS: instâncias de contêineres do Amazon ECS
- Contêineres do EKS: nós de processamento do EKS
- Funções: configuração de funções do Lambda
- Controle do estado do processo para sua instância do EC2

Vídeos relacionados:

- Amazon EC2 foundations (CMP211-R2)
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (CMP202-R1) (Computação melhor, mais rápida e com custo mais baixo: Otimização de custos com o EC2 (CMP202-R1))
- Deliver high performance ML inference with AWS Inferentia (CMP324-R1) (Entregar inferência de ML de alta performance com o AWS Inferentia (CMP324-R1))
- Optimize performance and cost for your AWS compute (CMP323-R1) (Otimizar a performance e os custos de sua computação da AWS (CMP323-R1))
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- How to choose compute option for startups (Como escolher uma opção de computação para startups)
- Optimize performance and cost for your AWS compute (CMP323-R1) (Otimizar a performance e os custos de sua computação da AWS (CMP323-R1))

Exemplos relacionados:

- Rightsizing with Compute Optimizer and Memory utilization enabled (Dimensionamento correto com o AWS Compute Optimizer e utilização da memória ativada)
- Código para demonstração do AWS Compute Optimizer

PERF02-BP05 Usar a elasticidade dos recursos disponíveis

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 as métricas relacionadas à computação, uma workload pode responder automaticamente a mudanças e utilizar um conjunto ideal de recursos para atingir sua meta.

A combinação ideal entre oferta e demanda leva ao menor custo para uma carga de trabalho, mas você também precisa se planejar para que exista oferta suficiente a fim de permitir tempo de

provisionamento e falhas de recursos individuais. A demanda pode ser fixa ou variável, exigindo métricas e automação para garantir que o gerenciamento não se torne um custo pesado e desproporcionalmente grande.

Na AWS, é possível usar várias abordagens diferentes para corresponder o suprimento com a demanda. O whitepaper Pilar Otimização de custos descreve como usar as seguintes abordagens de custo:

- Abordagem baseada em demanda
- Abordagem baseada em buffer
- Abordagem baseada em tempo

É necessário garantir que as implantações de carga de trabalho possam lidar com eventos de expansão e redução da escala. Crie cenários de teste para eventos de redução da escala a fim de garantir que a carga de trabalho se comporte conforme o esperado.

Antipadrões comuns:

- Reaja a alarmes aumentando a capacidade manualmente.
- Você deixa a capacidade aumentada após um evento de escalabilidade, em vez de reduzir novamente.

Benefícios do estabelecimento desta prática recomendada: A configuração e os testes da elasticidade da workload ajudam a fazer economia, manter as referências da performance e melhorar a confiabilidade à medida que o tráfego muda. A maior parte das instâncias que não são de produção deve ser interrompida quando não estiver em uso. Embora seja possível desligar manualmente instâncias não utilizadas, isso é impraticável em escalas maiores. Você também pode aproveitar a elasticidade baseada em volume, o que permite otimizar a performance e o custo, aumentando automaticamente o número de instâncias de computação durante picos de demanda e diminuindo a capacidade quando a demanda é reduzida.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

Aproveitar a elasticidade: a elasticidade corresponde ao suprimento de recursos que você tem em relação à demanda por esses recursos. Instâncias, contêineres e funções oferecem mecanismos para elasticidade, seja em combinação com a escalabilidade automática ou como um recurso do

serviço. Use a elasticidade em sua arquitetura para garantir que haja capacidade suficiente para atender aos requisitos de performance em todas as escalas de uso. Certifique-se de que as métricas para aumentar ou reduzir recursos elásticos sejam validadas em relação ao tipo de carga de trabalho que está sendo implantada. Se você estiver implantando uma aplicação de transcodificação de vídeo, espera-se que a utilização da CPU seja de 100%, e essa não deve ser sua métrica principal. Como alternativa, você pode medir em relação ao comprimento da fila de trabalhos de transcodificação aguardando para escalar seus tipos de instância. É necessário garantir que as implantações de carga de trabalho possam lidar com eventos de expansão e redução da escala. Reduzir os componentes da carga de trabalho com segurança é tão essencial quanto aumentar a escala de recursos quando a demanda exige. Crie cenários de teste para eventos de redução da escala a fim de garantir que a carga de trabalho se comporte conforme o esperado.

Recursos

Documentos relacionados:

- Computação em nuvem com a AWS
- Tipos de instância do EC2
- Contêineres do ECS: instâncias de contêineres do Amazon ECS
- Contêineres do EKS: nós de processamento do EKS
- Funções: configuração de funções do Lambda
- Controle do estado do processo para sua instância do EC2

Vídeos relacionados:

- Amazon EC2 foundations (CMP211-R2)
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (CMP202-R1) (Computação melhor, mais rápida e com custo mais baixo: Otimização de custos com o EC2 (CMP202-R1))
- Deliver high performance ML inference with AWS Inferentia (CMP324-R1) (Entregar inferência de ML de alta performance com o AWS Inferentia (CMP324-R1))
- Optimize performance and cost for your AWS compute (CMP323-R1) (Otimizar a performance e os custos de sua computação da AWS (CMP323-R1))
- Powering next-gen Amazon EC2: Deep dive into the Nitro system

Exemplos relacionados:

- Exemplos de grupos do Amazon EC2 Auto Scaling
- Tutoriais do Amazon EFS

PERF02-BP06 Reavaliar as necessidades de computação com base em 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.

Antipadrões comuns:

- Você só monitora métricas no nível do sistema para obter informações sobre sua carga de trabalho.
- Você arquiteta suas necessidades de computação para os requisitos de pico de carga de trabalho.
- Você superdimensiona uma solução de computação para atender aos requisitos de escalabilidade de performance, quando uma nova solução de computação corresponderia às características da sua workload.

Benefícios do estabelecimento desta prática recomendada: Para otimizar a performance e a utilização de recursos, você precisa de uma visão operacional unificada, dados granulares em tempo real e uma referência histórica. Você pode criar painéis automáticos para visualizar esses dados e executar matemática de métricas para obter informações operacionais e de utilização.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

Usar uma abordagem direcionada a dados para otimizar os recursos: para obter a máxima performance e eficiência, use os dados coletados da workload ao longo do tempo para ajustar e otimizar seus recursos. Analise as tendências no uso dos recursos atuais da sua carga de trabalho e determine em que você pode fazer alterações para atender melhor às necessidades da sua carga de trabalho. A performance do sistema cai quando os recursos estão sendo comprometidos excessivamente, já a subutilização de recursos leva a um uso menos eficiente e maiores custos dos mesmos.

Recursos

Documentos relacionados:

- Computação em nuvem com a AWS
- AWS Compute Optimizer
- Computação em nuvem com a AWS
- Tipos de instância do EC2
- Contêineres do ECS: instâncias de contêineres do Amazon ECS
- Contêineres do EKS: nós de processamento do EKS
- Funções: configuração de funções do Lambda
- Controle do estado do processo para sua instância do EC2

Vídeos relacionados:

- Amazon EC2 foundations (CMP211-R2)
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (CMP202-R1) (Computação melhor, mais rápida e com custo mais baixo: Otimização de custos com o EC2 (CMP202-R1))
- Deliver high performance ML inference with AWS Inferentia (CMP324-R1) (Entregar inferência de ML de alta performance com o AWS Inferentia (CMP324-R1))
- Optimize performance and cost for your AWS compute (CMP323-R1) (Otimizar a performance e os custos de sua computação da AWS (CMP323-R1))
- Powering next-gen Amazon EC2: Deep dive into the Nitro system

Exemplos relacionados:

- Rightsizing with Compute Optimizer and Memory utilization enabled (Dimensionamento correto com o AWS Compute Optimizer e utilização da memória ativada)
- Código para demonstração do AWS Compute Optimizer

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.

Práticas recomendadas

- PERF03-BP01 Compreender as características e os requisitos de armazenamento
- PERF03-BP02 Avaliar as opções de configuração disponíveis
- PERF03-BP03 Tomar decisões com base em padrões de acesso e métricas

PERF03-BP01 Compreender as características e os requisitos de armazenamento

Identifique e documente as necessidades de armazenamento de workloads e defina as características de armazenamento de cada local. Exemplos de características de armazenamento incluem: acesso compartilhável, tamanho de arquivo, taxa de crescimento, throughput, IOPS, latência, padrões de acesso e persistência dos dados. Use essas características para avaliar se os serviços de armazenamento de blocos, arquivos, objetos ou instâncias são a solução mais eficiente para suas necessidades de armazenamento.

Resultado desejado: identifique e documente os requisitos para cada armazenamento e avalie as soluções de armazenamento disponíveis. Com base nas principais características de armazenamento, sua equipe vai entender como os serviços de armazenamento selecionados vão beneficiar o desempenho de sua workload. Os principais critérios incluem os padrões de acesso aos dados, a taxa de crescimento, as necessidades de escalabilidade e os requisitos de latência.

Antipadrões comuns:

- Você só usa um tipo de armazenamento, como o Amazon Elastic Block Store (Amazon EBS), para todas as workloads.
- Você pressupõe que todas as cargas de trabalho têm requisitos semelhantes de performance de acesso ao armazenamento.

Benefícios do estabelecimento desta prática recomendada: selecionar a solução de armazenamento com base nas características identificadas e necessárias vai ajudar você a melhorar a performance de suas workloads, reduzir os custos e diminuir os esforços operacionais para manter a workload. A performance de sua workload vai se beneficiar da solução, da configuração e do local do serviço de armazenamento.

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

Orientação de implementação

Identifique as métricas de performance de armazenamento mais importantes da sua carga de trabalho e implemente melhorias como parte de uma abordagem impulsionada por dados, usando testes comparativos ou testes de carga. Use esses dados para identificar em que ponto sua solução de armazenamento é restrita e examinar as opções de configuração que possam melhorar a solução. Determine a taxa de crescimento esperada para sua carga de trabalho e escolha uma solução de armazenamento que atenda a essas taxas. Pesquise as ofertas de armazenamento da AWS para determinar a solução de armazenamento correta para as necessidades de sua workload. O provisionamento de soluções de armazenamento na AWS aumenta a oportunidade para você testar ofertas de armazenamento e determinar se são adequadas para as necessidades de sua workload.

Serviço da AWS	Características principais	Casos de uso comum
Amazon S3	99,999999999% de durabilidade, crescimento ilimitado, acessível de qualquer lugar, vários modelos de custo baseados em acesso e resiliência.	Dados de aplicações nativas de nuvem, arquivamento de dados, backups, análises, data lakes, hospedagem de site estático, dados de IoT.
Amazon S3 Glacier	Latência de segundos a horas, crescimento ilimitado, menor custo, armazenamento de longo prazo.	Arquivamento de dados, arquivos de mídia, retenção de backup de longo prazo.
Amazon EBS	O tamanho do armazenamento requer gerenciamento e monitoramento, baixa latência, armazenamento persistente, 99,8% a 99,9% de durabilidade, a maioria dos tipos de volume só podem ser acessados de uma instância do EC2.	Aplicações COTS, aplicações com uso intenso de E/S, bancos de dados relacionais e NoSQL, backup e recuperação.

Serviço da AWS	Características principais	Casos de uso comum
Armazenamento de instâncias do EC2	Tamanho de armazenamento predeterminado, menor latência, sem persistência, acessível somente de uma instância do EC2.	Aplicações COTS, aplicações com uso intenso de E/S, datastore na memória.
Amazon EFS	99,999999999% de durabilidade, crescimento ilimitado, acessível de vários serviços de computação.	Aplicações modernas compartilhando arquivos entre vários serviços de computação, armazenamento de arquivos para dimensionamento de sistemas de gerenciamento de conteúdo.
Amazon FSx	Compatível com quatro sistemas de arquivos (NetApp, OpenZFS, Windows File Server e Amazon FSx for Lustre), o armazenamento disponível varia de acordo com o sistema de arquivos, acessível de vários serviços de computação.	Workloads nativas de nuvem, expansão na nuvem privada, workloads migradas que exigem um sistema de arquivos específico, VMC, sistemas de ERP, backups e armazenamento de arquivos on-premises.
Família Snow	Dispositivos portáteis, criptografia de 256 bits, endpoint NFS, computação integrada, TBs de armazenamento.	Migração de dados para a nuvem, armazenamento e computação em condições on-premises extremas, recuperação de desastres, coleta de dados remota.

Serviço da AWS	Características principais	Casos de uso comum
AWS Storage Gateway	Oferece acesso on-premises de baixa latência ao armazenamento com backup na nuvem, cache on-premises totalmente gerenciado.	Migrações de dados on-premises para a nuvem, preenchimento de data lakes na nuvem usando origens on-premises, compartilhamento de dados modernizado.

Etapas da implementação:

1. use testes de carga ou benchmarking para coletar as principais características de suas necessidades de armazenamento. As principais características incluem:
   a. Compartilhável (quais componentes acessam esse armazenamento)
   b. Taxa de crescimento
   c. Taxa de transferência
   d. Latência
   e. Tamanho de E/S
   f. Durabilidade
   g. Padrões de acesso (leituras vs. gravações, frequência, com picos ou consistente)
2. Identifique o tipo de solução de armazenamento compatível com as características do seu armazenamento.
   a. Amazon S3 é um serviço de armazenamento de objetos com escalabilidade ilimitada, alta disponibilidade e várias opções de acessibilidade. A transferência e o acesso a objetos dentro e fora do Amazon S3 podem usar um serviço, como Aceleração de Transferências ou Pontos de Acesso , para oferecer suporte ao seu local, necessidades de segurança e padrões de acesso. Use a ferramenta de recomendações do diretrizes de performance do Amazon S3 para ajudar você a otimizar sua configuração do Amazon S3 e atender às necessidades de performance da workload.
   b. Amazon S3 Glacier é uma classe de armazenamento do Amazon S3 desenvolvida para arquivamento de dados. Você pode escolher entre três soluções de arquivamento com acesso que varia de milissegundos até 5 a 12 horas com diversas opções de custo e segurança. O Amazon S3 Glacier pode ajudar você a cumprir os requisitos de performance ao

implementar um ciclo de vida de dados que ofereça suporte aos seus requisitos de negócios e características de dados.

c. Amazon Elastic Block Store (Amazon EBS) é um serviço de armazenamento de blocos de alta performance projetado para o Amazon Elastic Compute Cloud (Amazon EC2). Você pode escolher entre soluções baseadas em SSD ou HDD com características diferentes que priorizam IOPS ou throughput. Os volumes do EBS são adequados para workloads de alta performance, armazenamento primário para sistemas de arquivos, bancos de dados ou aplicações que só podem acessar sistemas de estágio associado.

d. Armazenamento de instâncias do Amazon EC2 é semelhante ao Amazon EBS já que se associa a uma instância do Amazon EC2, mas o armazenamento de instância é apenas um armazenamento temporário que, idealmente, deve ser usado como buffer, cache ou outro conteúdo temporário. Não é possível desassociar um armazenamento de instância e todos os dados serão perdidos se a instância for encerrada. Armazenamentos de instâncias podem ser usados para casos de uso de alta performance de E/S e baixa latência em que os dados não precisam persistir.

e. Amazon Elastic File System (Amazon EFS) é um sistema de arquivos montável que pode ser acessado por diversos tipos de soluções de computação. O Amazon EFS aumenta e reduz automaticamente o armazenamento e sua performance é otimizada para oferecer latências baixas de maneira consistente. O EFS tem dois modos de configuração de performance: Propósito geral e E/S Máx. Propósito geral tem latência de leitura inferior a milissegundo e latência de gravação que nunca chega a 10 milissegundos. O recurso E/S Máx. oferece suporte a milhares de instâncias de computação que exigem um sistema de arquivos compartilhado. O Amazon EFS oferece suporte a dois modos de throughput: expansão e provisionada. Uma workload que tem um padrão de acesso com picos vai se beneficiar do modo de throughput de expansão, enquanto uma workload consistentemente alta tem melhor performance com o modo de throughput provisionada.

f. Amazon FSx se baseia nas soluções de computação mais recentes da AWS para oferecer suporte a quatro sistemas de arquivos comumente usados: NetApp ONTAP, OpenZFS, Windows File Server e Lustre. A latência, throughput e IOPS do Amazon FSx variam de acordo com o sistema de arquivos e devem ser consideradas ao selecionar o sistema de arquivos certo para as necessidades de sua workload.

g. AWS Snow Family consiste em dispositivos de armazenamento e computação que oferecem suporte à migração de dados online e offline para a nuvem, além de armazenamento de dados e computação on-premises. Os dispositivos AWS Snow oferecem suporte à coleta de grandes quantidades de dados on-premises, processamento desses dados e movimentação desses

dados para a nuvem. Há diversas práticas recomendadas e documentadas sobre performance no que se refere a número de arquivos, tamanhos de arquivos e compressão.

h. AWS Storage Gateway oferece a aplicações on-premises acesso ao armazenamento baseado em nuvem. O AWS Storage Gateway é compatível com vários serviços de armazenamento em nuvem, incluindo Amazon S3, Amazon S3 Glacier, Amazon FSx e Amazon EBS. Ele oferece suporte a diversos protocolos, como iSCSI, SMB e NFS. Oferece performance de baixa latência ao armazenar em cache os dados acessados com frequência on-premises e só envia dados alterados e comprimidos à AWS.

3. Depois de experimentar a nova solução de armazenamento e identificar a configuração ideal, planeje a migração e valide as métricas de performance. Esse processo é contínuo e deve ser reavaliado quando houver mudança em características importantes ou quando os serviços e as opções disponíveis mudarem.

Nível de esforço do plano de implementação: Se uma workload estiver sendo movida de uma solução de armazenamento para outra, poderá haver um nível moderado de esforço envolvido na refatoração da aplicação.

Recursos

Documentos relacionados:

- Tipos de volume do Amazon EBS
- Armazenamento do Amazon EC2
- Amazon EFS: performance do Amazon EFS
- Performance do Amazon FSx para Lustre
- Performance do Amazon FSx for Windows File Server
- Performance do Amazon FSx for NetApp ONTAP
- Performance do Amazon FSx for OpenZFS
- Amazon S3 Glacier: documentação do Amazon S3 Glacier
- Amazon S3: considerações sobre performance e taxa de solicitação
- Armazenamento na nuvem com a AWS
- AWS Snow Family
- Características de E/S do EBS

Vídeos relacionados:

- Deep dive on Amazon EBS (STG303-R1)
- Optimize your storage performance with Amazon S3 (STG343)

Exemplos relacionados:

- Driver CSI do Amazon EFS
- Driver CSI do Amazon EBS
- Utilitários do Amazon EFS
- Escalabilidade automática do Amazon EBS
- Exemplos do Amazon S3
- Driver CSI (Interface de armazenamento de contêiner) do Amazon FSx for Lustre

PERF03-BP02 Avaliar 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.

Amazon EBS disponibiliza várias opções que permitem otimizar a performance do armazenamento e o custo para a sua carga de trabalho. Essas opções estão divididas em duas categorias principais: armazenamento baseado em SSD para cargas de trabalho transacionais, como bancos de dados e volumes de inicialização (a performance depende principalmente de IOPS), e armazenamento baseado em HDD para cargas de trabalho com uso intenso de throughput, como MapReduce e processamento de logs (a performance depende principalmente de MB/s).

Volumes baseados em SSD incluem SSDs com a mais alta performance de IOPS provisionadas para workloads transacionais sensíveis à latência e para SSDs de uso geral, que equilibram preço e performance para uma ampla variedade de dados transacionais.

A aceleração de transferência do Amazon S3 permite a transferência rápida de arquivos em longas distâncias entre o cliente e o seu bucket do S3. A aceleração de transferência utiliza pontos de presença globalmente distribuídos do Amazon CloudFront para rotear dados por um caminho de rede otimizado. Para uma carga de trabalho em um bucket do S3 com muitas solicitações GET, use

o Amazon S3 com o CloudFront. Ao fazer upload de arquivos grandes, use uploads em várias partes, carregando-as de uma só vez para ajudar a maximizar a taxa de transferência de rede.

O Amazon Elastic File System (Amazon EFS) fornece um sistema elástico de arquivos NFS simples, escalável e totalmente gerenciado para uso com serviços da Nuvem AWS e recursos on-premises. Para compatibilidade com uma grande variedade de workloads de armazenamento na nuvem, o Amazon EFS oferece dois modos de performance: o modo de performance de uso geral e o modo de performance máxima de E/S. Também há dois modos de throughput a escolher para o sistema de arquivos: throughput com expansão e throughput provisionado. Para determinar quais configurações usar para sua carga de trabalho, consulte o Guia do usuário do Amazon EFS.

Amazon FSx fornece quatro sistemas de arquivos para escolher: o Amazon FSx for Windows File Server para workloads empresariais. O Amazon FSx for Lustre para workloads de alta performance. O Amazon FSx for NetApp ONTAP para o sistema de arquivos ONTAP popular para NetApps e o Amazon FSx for OpenZFS para servidores de arquivos baseados em Linux. O FSx é baseado em SSD e é projetado para fornecer performance rápida, previsível, dimensionável e consistente. O sistema de arquivos do Amazon FSx fornece altas velocidades de leitura e gravação sustentáveis e acesso consistente a dados de baixa latência. Você pode escolher o nível de throughput necessário para atender às necessidades de sua carga de trabalho.

Antipadrões comuns:

- Você só usa um tipo de armazenamento, como o Amazon EBS, para todas as workloads.
- Você usa as IOPS provisionadas para todas as workloads sem testes reais em todos os níveis de armazenamento.
- Você pressupõe que todas as cargas de trabalho têm requisitos semelhantes de performance de acesso ao armazenamento.

Benefícios do estabelecimento desta prática recomendada: Avaliar todas as opções de serviço de armazenamento pode reduzir o custo da infraestrutura e o esforço necessário para manter suas cargas de trabalho. Isso pode acelerar potencialmente seu tempo de entrada no mercado para a implantação de novos serviços e recursos.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

Determinar as características de armazenamento: ao avaliar a solução de armazenamento, determine as características de armazenamento de que você precisa, como a habilidade de

compartilhar, o tamanho dos arquivos, o tamanho do cache, a latência, o throughput e a persistência dos dados. Corresponda os requisitos ao serviço da AWS mais adequado às suas necessidades.

Recursos

Documentos relacionados:

- Armazenamento na nuvem com a AWS
- Tipos de volume do Amazon EBS
- Armazenamento do Amazon EC2
- Amazon EFS: performance do Amazon EFS
- Performance do Amazon FSx for Lustre
- Performance do Amazon FSx for Windows File Server
- Amazon Glacier: documentação do Amazon Glacier
- Amazon S3: considerações sobre performance e taxa de solicitações
- Armazenamento na nuvem com a AWS
- Armazenamento na nuvem com a AWS
- Características de E/S do EBS

Vídeos relacionados:

- Deep dive on Amazon EBS (STG303-R1)
- Optimize your storage performance with Amazon S3 (STG343)

Exemplos relacionados:

- Amazon EFS CSI Driver (Driver CSI do Amazon EFS)
- Amazon EBS CSI Driver (Driver CSI do Amazon EBS)
- Amazon EFS Utilities (Utilitários do EFS)
- Amazon EBS Autoscale (Escalabilidade automática do Amazon EBS)
- Exemplos do Amazon S3

PERF03-BP03 Tomar decisões com base em padrões de acesso e 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 armazenamento 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.

A maneira como você acessa os dados afeta a performance da solução de armazenamento. Selecione a solução de armazenamento mais alinhada a seus padrões de acesso ou considere a possibilidade de alterar seus padrões de acesso para alinhamento com a solução de armazenamento a fim de maximizar a performance.

A criação de uma matriz RAID 0 permite atingir um nível maior de performance para um sistema de arquivos em relação ao que é possível provisionar em um único volume. Avalie a possibilidade de usar RAID 0 quando a performance de E/S for mais importante que a tolerância a falhas. Por exemplo, você poderia usá-la com um banco de dados muito utilizado e no qual a replicação de dados já esteja configurada separadamente.

Selecione as métricas de armazenamento adequadas para sua carga de trabalho em todas as opções de armazenamento consumidas para a carga de trabalho. Ao utilizar sistemas de arquivos que usam créditos de expansão, crie alarmes para ser informado quando você estiver se aproximando dos limites desses créditos. É necessário criar painéis de armazenamento para mostrar a integridade geral do armazenamento da carga de trabalho.

Para sistemas de armazenamento com tamanho fixo, como o Amazon EBS ou o Amazon FSx, certifique-se de estar monitorando a quantidade de armazenamento usada em comparação com o tamanho geral do armazenamento e, se possível, crie automação para aumentar o tamanho do armazenamento ao atingir um limite

Antipadrões comuns:

- Você pressupõe que a performance do armazenamento seja adequada se os clientes não estiverem reclamando.
- Você usa apenas um nível de armazenamento, supondo que todas as cargas de trabalho se encaixem nesse nível.

Benefícios do estabelecimento desta prática recomendada: Você precisa de uma visão operacional unificada, dados granulares em tempo real e uma referência histórica para otimizar a performance

e a utilização de recursos. É possível criar painéis e dados automáticos com granularidade de um segundo para executar matemática de métricas nos dados e para fornecer insights operacionais e de utilização para as suas necessidades de armazenamento.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

Otimizar o uso do armazenamento e os padrões de acesso: escolha sistemas de armazenamento com base nos padrões de acesso de sua workload e nas características das opções de armazenamento disponíveis. Determine o melhor local para armazenar os dados que permitirão que você atenda a seus requisitos enquanto reduz a sobrecarga. Use otimizações de performance e padrões de acesso ao configurar e interagir com dados conforme as características de seu armazenamento (por exemplo, remover volumes ou particionar os dados).

Selecionar as métricas adequadas para as opções de armazenamento: selecione as métricas de armazenamento adequadas para a workload. Cada opção de armazenamento oferece várias métricas para acompanhar a performance da carga de trabalho ao longo do tempo. Verifique se você está realizando medições em relação a métricas de expansão de armazenamento (por exemplo, monitoramento de créditos de expansão para o Amazon EFS). Para sistemas de armazenamento de tamanho fixo, como o Amazon Elastic Block Store ou o Amazon FSx, verifique se você está monitorando a quantidade de armazenamento usada em comparação com o tamanho geral do armazenamento. Crie automação quando possível para aumentar o tamanho do armazenamento ao atingir um limite.

Monitorar métricas: o Amazon CloudWatch pode coletar métricas entre os recursos da sua arquitetura. Você também pode coletar e publicar métricas personalizadas para descobrir métricas de negócio ou derivadas. Use o CloudWatch ou soluções de terceiros para definir alarmes que indiquem quando os limites são violados.

Recursos

Documentos relacionados:

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

- Performance do Amazon FSx for Windows File Server
- Amazon Glacier: documentação do Amazon Glacier
- Amazon S3: considerações sobre performance e taxa de solicitações
- Armazenamento na nuvem com a AWS
- Características de E/S do EBS
- Monitoring and understanding Amazon EBS performance using Amazon CloudWatch (Monitoramento e compreensão da performance do Amazon EBS usando o Amazon CloudWatch)

Vídeos relacionados:

- Deep dive on Amazon EBS (STG303-R1)
- Optimize your storage performance with Amazon S3 (STG343)

Exemplos relacionados:

- Amazon EFS CSI Driver (Driver CSI do Amazon EFS)
- Amazon EBS CSI Driver (Driver CSI do Amazon EBS)
- Amazon EFS Utilities (Utilitários do EFS)
- Amazon EBS Autoscale (Escalabilidade automática do Amazon EBS)
- Exemplos do Amazon S3

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. A seleção da solução e dos recursos de banco de dados incorretos para um sistema pode levar a uma menor performance do sistema.

Práticas recomendadas

- PERF04-BP01 Compreender as características dos dados
- PERF04-BP02 Avaliar as opções disponíveis
- PERF04-BP03 Coletar e registrar métricas de performance do banco de dados
- PERF04-BP04 Escolher armazenamento de dados com base nos padrões de acesso

- PERF04-BP05 Otimizar o armazenamento de dados com base nas métricas e nos padrões de acesso

PERF04-BP01 Compreender as características dos dados

Escolha soluções para o gerenciamento dos seus dados que correspondam de forma ideal às características, aos padrões de acesso e aos requisitos dos conjuntos de dados de sua workload. Ao selecionar e implementar uma solução de gerenciamento de dados, verifique se as características de consultas, de escalabilidade e de armazenamento são compatíveis com os requisitos dos dados da workload. Saiba como as várias opções de bancos de dados correspondem aos seus modelos de dados e quais opções de configuração são mais adequadas para seu caso de uso.

A AWS fornece vários mecanismos de banco de dados, incluindo bancos de dados relacionais, de chave-valor, de documentos, na memória, de grafos, de séries temporais e de ledger. Cada solução de gerenciamento de dados tem opções e configurações disponíveis para compatibilidade com seus casos de uso e modelos de dados. Sua workload deve poder usar várias soluções de banco de dados diferentes, baseadas nas características dos dados. Ao selecionar as melhores soluções de banco de dados para um problema específico, você pode se libertar de bancos de dados monolíticos, com a abordagem de tamanho único que é restritiva, e focar no gerenciamento de dados para atender às necessidades dos seus clientes.

Resultado desejado: As características dos dados da workload são documentadas com detalhes suficientes para facilitar a seleção e a configuração de soluções de banco de dados compatíveis e para fornecer insight das possíveis alternativas.

Antipadrões comuns:

- Não considerar maneiras para segmentar grandes conjuntos de dados em coleções de dados menores que têm características semelhantes, o que resulta na perda das oportunidades de usar bancos de dados com propósito específico que correspondem melhor às características dos dados e do crescimento.
- Não identificar os padrões de acesso aos dados no início, o que resulta em retrabalho caro e complexo posteriormente.
- Limitar o crescimento usando estratégias de armazenamento de dados que não são dimensionáveis tão rapidamente quanto necessário.
- Escolher um tipo de banco de dados para todas as workloads.
- Fixar-se em uma única solução de banco de dados porque há experiência e conhecimento internos de um tipo específico de solução de banco de dados.

- Manter uma solução de banco de dados porque ela funciona bem em um ambiente on-premises.

Benefícios do estabelecimento desta prática recomendada: Familiarize-se com todas as soluções de banco de dados da AWS para poder determinar a solução de banco de dados correta para as diversas workloads. Depois de selecionar a solução de banco de dados adequada para a sua workload, você poderá experimentar rapidamente cada uma dessas ofertas de banco de dados para determinar se elas continuam atendendo às necessidades da sua workload.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

- A economia de custo possível pode não ser identificada.
- Os dados podem não estar protegidos no nível necessário.
- O acesso aos dados e a performance do armazenamento podem não ser ideais.

Orientações para a implementação

Defina as características dos dados e os padrões de acesso da workload. Analise todas as soluções de banco de dados disponíveis para identificar a solução que é compatível com as necessidades dos seus dados. Em uma determinada workload, vários bancos de dados podem ser selecionados. Avalie cada serviço ou grupo de serviços e analise-os individualmente. Se soluções alternativas possíveis de gerenciamento de dados forem identificadas para todos ou para parte dos dados, experimente com implementações alternativas que possam desvendar benefícios de custo, segurança, performance e confiabilidade. Atualize a documentação existente caso uma nova abordagem de gerenciamento de dados seja adotada.

Tipo	Serviços da AWS	Características principais	Casos de uso comuns
Relacional	Amazon RDS, Amazon Aurora	Integridade referencial, transações ACID, esquema para gravação	ERP, CRM, software comercial pronto para uso
Chave-valor	Amazon DynamoDB	Throughput alto, baixa latência, escalabilidade quase infinita	Carrinhos de compras (comércio eletrônico), catálogos de

Tipo	Serviços da AWS	Características principais	Casos de uso comuns
			produtos, aplicações de chat
Documentos	Amazon DocumentDB	Armazenar documentos JSON e consultar por qualquer atributo	Gerenciamento de conteúdo (CMS), perfis de clientes, aplicações móveis
Na memória	Amazon ElastiCache, Amazon MemoryDB	Latência de microssegundos	Armazenamento em cache, placares de jogos
Bancos de dados	Amazon Neptune	Dados altamente relacionais em que os relacionamentos entre os dados têm significado	Redes sociais, mecanismos de personalização, detecção de fraudes
Séries temporais	Amazon Timestream	Dados em que a dimensão primária é o tempo	DevOps, IoT, monitoramento
Coluna ampla	Amazon Keyspaces	Workloads do Cassandra.	Manutenção de equipamentos industriais, otimização de rotas
Ledger	Amazon QLDB	Ledger de alterações imutável e verificável de forma criptográfica	Sistemas de registro, saúde, cadeias de suprimentos, instituições financeiras

Etapas da implementação

1. Como os dados estão estruturados? (por exemplo, não estruturados, chave-valor, semiestruturados, relacionais)
    a. Se os dados estiverem estruturados, considere um armazenamento de objetos, como o Amazon S3 , ou um banco de dados NoSQL, como o Amazon DocumentDB.
    b. Para dados de chave-valor, considere o DynamoDB, o ElastiCache for Redis ou o MemoryDB.
    c. Se os dados tiverem uma estrutura relacional, qual nível de integridade referencial é necessário?
        i. Para restrições de chave estrangeira, bancos de dados relacionais, como o Amazon RDS e Aurora , podem fornecer esse nível de integridade.
        ii. Normalmente, em um modelo de dados NoSQL, você desnormalizaria os dados em um único documento ou coleção de documentos para serem recuperados em uma única solicitação e não fazendo junção entre documentos ou tabelas;
2. A conformidade com as ACID (atomicidade, consistência, isolamento, durabilidade) é necessária?
    a. Se as propriedades ACID associadas aos bancos de dados relacionais forem necessárias, considere um banco de dados relacional, como o Amazon RDS e o Aurora.
3. Qual é o modelo de consistência necessário?
    a. Se a sua aplicação puder tolerar consistência eventual, considere uma implementação NoSQL. Analise as outras características para ajudar a escolher qual banco de dados NoSQL é mais adequado.
    b. Se for necessária forte consistência, use leituras altamente consistentes com o DynamoDB ou um banco de dados relacional, como o Amazon RDS.
4. Quais formatos de consulta e resultado são compatíveis? (por exemplo, SQL, CSV, Parque, Avro, JSON etc.)
5. Quais tipos de dados, tamanhos de campos e quantidades gerais estão presentes? (por exemplo, texto, numérico, espacial, séries temporais calculadas, binário ou blob, documento)
6. Como as necessidades de armazenamento serão alteradas ao longo do tempo? Como isso afeta a escalabilidade?
    a. Bancos de dados de tecnologia sem servidor, como o DynamoDB e Amazon Quantum Ledger Database , escalarão dinamicamente até quase armazenamento ilimitado.
    b. Os bancos de dados relacionais têm limites superiores em armazenamento provisionado e devem ser particionados horizontalmente por meio de mecanismos, como fragmentação, quando atingem esses limites.

7. Qual é a proporção de consultas de leitura em relação a consultas de gravação? O armazenamento em cache melhoraria a performance?
    a. Workloads de leitura pesada podem se beneficiar de uma camada de armazenamento em cache, esse pode ser o ElastiCache ou o DAX, se o banco de dados for o DynamoDB.
    b. As leituras também podem ser descarregadas em réplicas de leitura com bancos de dados relacionais, como o Amazon RDS.
8. O armazenamento e a modificação (OLTP – Processamento de transações on-line) ou a recuperação e a geração de relatórios (OLAP – Processamento analítico on-line) têm uma prioridade mais alta.
    a. Para processamento transacional de throughput alto, considere um banco de dados NoSQL, como o DynamoDB ou o Amazon DocumentDB.
    b. Para consultas de análise, considere um banco de dados em colunas, como o Amazon Redshift , ou exporte os dados para o Amazon S3 e execute análises usando o Athena ou o QuickSight.
9. Qual é o grau de confidencialidade desses dados e qual nível de proteção e criptografia eles precisam?
    a. Todos os mecanismos do Amazon RDS e do Aurora são compatíveis com a criptografia de dados em repouso usando o AWS KMS. O Microsoft SQL Server e o Oracle também são compatíveis com a Transparent Data Encryption (TDE – Criptografia transparente de dados) ao usar o Amazon RDS.
    b. Para o DynamoDB, use controle de acesso refinado com o IAM para controlar quem tem acesso a quais dados no nível principal.
10. Qual nível de durabilidade os dados exigem?
    a. O Aurora replica automaticamente os dados entre três zonas de disponibilidade em uma região, o que significa que seus dados são altamente duráveis com menos chance de perda de dados.
    b. O DynamoDB é automaticamente replicado entre várias zonas de disponibilidade, fornecendo alta disponibilidade e durabilidade dos dados.
    c. O Amazon S3 fornece 11 noves de durabilidade. Muitos serviços de banco de dados, como o Amazon RDS e o DynamoDB, são compatíveis com a exportação de dados para o Amazon S3 para retenção de longo prazo e arquivamento.
11. Os requisitos do objetivo de tempo de recuperação (RTO) ou do objetivo de ponto de recuperação (RPO) influenciam a solução?
    a. O Amazon RDS, o Aurora, o DynamoDB, o Amazon DocumentDB e o Neptune são compatíveis com a recuperação pontual e o backup e a recuperação sob demanda.

b. Para requisitos de alta disponibilidade, as tabelas do DynamoDB podem ser replicadas globalmente usando o recurso tabelas globais, e os clusters do Aurora podem ser replicados entre várias regiões usando o recurso banco de dados global. Além disso, os buckets do S3 podem ser replicados entre Regiões da AWS usando a replicação entre regiões.

12. Você quer se livrar de mecanismos de bancos de dados comerciais/custos de licenças?

a. Considere os mecanismos de código aberto, como o PostgreSQL e o MySQL no Amazon RDS ou no Aurora

b. Utilize o AWS DMS e o AWS SCT para executar migrações de mecanismos de bancos de dados comerciais para código aberto

13. Qual a expectativa operacional para o banco de dados? A mudança para serviços gerenciados é uma preocupação principal?

a. Utilizar o Amazon RDS em vez do Amazon EC2 e o DynamoDB ou o Amazon DocumentDB em vez de um host automático de um banco de dados NoSQL pode reduzir a sobrecarga operacional.

14. Como o banco de dados é acessado atualmente? É acessado apenas por aplicação ou há usuários de inteligência de negócios (BI) e outras aplicações prontas para uso conectadas?

a. Se você tiver dependências de ferramentas externas, poderá ser necessário manter a compatibilidade com os bancos de dados com os quais elas são compatíveis. O Amazon RDS é totalmente compatível com as diferentes versões de mecanismo aos quais oferece suporte, incluindo o Microsoft SQL Server, o Oracle, o MySQL e o PostgreSQL.

15. Veja a seguir uma lista de serviços dos gerenciamento de dados potenciais, e onde eles podem ser melhor utilizados:

a. Bancos de dados relacionais armazenam dados com esquemas e relacionamentos predefinidos entre eles. Esses bancos de dados são projetados para oferecer suporte a transações ACID (atomicidade, consistência, isolamento, durabilidade) e manter a integridade referencial e uma forte consistência de dados. Muitas aplicações tradicionais, planejamento de recursos empresariais (ERP), gerenciamento de relacionamentos com o cliente (CRM) e comércio eletrônico usam bancos de dados relacionais para armazenar seus dados. É possível executar muitos desses mecanismos de banco de dados no Amazon EC2 ou escolher um dos serviços gerenciados pela AWS de banco de dados: Amazon Aurora, Amazon RDSe aos Amazon Redshift.

b. Bancos de dados de chave/valor são otimizados para padrões de acesso comuns, normalmente visando armazenar e recuperar grandes volumes de dados. Esses bancos de dados fornecem tempos de resposta rápidos, mesmo sob volumes extremos de solicitações

simultâneas. Aplicações da web de alto tráfego, sistemas de comércio eletrônico e aplicações de jogos são os casos de uso habituais para bancos de dados de chave-valor. Na AWS, é possível utilizar o Amazon DynamoDB, um banco de dados totalmente gerenciado, multirregião, multimestre e durável com recursos incorporados de segurança, backup e restauração, além de armazenamento em cache na memória para aplicações na escala da Internet.

c. Os bancos de dados na memória são usados para aplicações que exigem acesso em tempo real aos dados, latência mais baixa e throughput mais alto. Ao armazenar dados diretamente na memória, esses bancos de dados fornecem latência de microssegundos às aplicações para as quais a latência de milissegundos não é suficiente. Você pode usar bancos de dados em memória para armazenamento de aplicativos em cache, gerenciamento de sessões, placares de jogos e aplicativos geoespaciais. Amazon ElastiCache é um datastore na memória totalmente gerenciado, compatível com o Redis ou Memcached. No caso de aplicações que também requerem durabilidade mais alta, Amazon MemoryDB for Redis oferece isso em combinação com um serviço durável de banco de dados na memória para performance ultrarrápida.

d. Um banco de dados de documentos é projetado para armazenar dados semiestruturados, como documentos semelhantes a JSON. Esses bancos de dados ajudam os desenvolvedores a criar e atualizar rapidamente aplicativos como gerenciamento de conteúdo, catálogos e perfis de usuário. Amazon DocumentDB é um serviço totalmente gerenciado de banco de dados de documentos rápido, escalável e altamente disponível compatível com cargas de trabalho do MongoDB.

e. Um armazenamento em colunas amplas é um tipo de banco de dados NoSQL. Ele usa tabelas, linhas e colunas, mas ao contrário de um banco de dados relacional, os nomes e o formato das colunas podem variar de linha para linha na mesma tabela. Normalmente, você vê um repositório de coluna ampla em aplicativos industriais de alta escala para manutenção de equipamentos, gerenciamento de frotas e otimização de rotas. Amazon Keyspaces (para Apache Cassandra) é um serviço escalável, gerenciado e altamente disponível de banco de dados de coluna ampla, compatível com Apache Cassandra.

f. Bancos de dados gráficos são para aplicativos que precisam navegar e consultar milhões de relações entre conjuntos de dados gráficos altamente conectados com latência de milissegundos em grande escala. Muitas empresas usam bancos de dados gráficos para detecção de fraudes, redes sociais e mecanismos de recomendação. Amazon Neptune é serviço totalmente gerenciado, rápido e confiável de banco de dados gráfico que facilita a criação e execução de aplicações que funcionam com conjuntos de dados altamente conectados.

g. Bancos de dados de séries temporais são eficientes para coletar, sintetizar e derivar insights de dados que mudam ao longo do tempo. Aplicativos de IoT, DevOps e telemetria industrial podem utilizar bancos de dados de séries temporais. Amazon Timestream é um serviço rápido, escalável e totalmente gerenciado de banco de dados de séries temporais para aplicativos operacionais e de IoT que facilita o armazenamento e a análise de trilhões de eventos por dia.

h. Bancos de dados de livro-razão fornecem uma autoridade centralizada e confiável para manter um registro escalável, imutável e criptograficamente verificável de transações para cada aplicativo. Vemos os bancos de dados de livro-razão empregados em sistemas de registro, cadeia de suprimentos, inscrições e até mesmo transações bancárias. Amazon Quantum Ledger Database (Amazon QLDB) é um banco de dados ledger totalmente gerenciado que fornece um log de transações transparente, imutável e criptograficamente verificável pertencente a uma autoridade confiável central. O Amazon QLDB monitora todas as alterações de dados do aplicativo e mantém um histórico completo e verificável das alterações ao longo do tempo.

Nível de esforço para o plano de implementação: Se uma workload for movida de uma solução de banco de dados para outra, poderá haver um nível alto de esforço envolvido na refatoração dos dados e da aplicação.

Recursos

Documentos relacionados:

- Bancos de dados em nuvem com a AWS
- Armazenamento em cache de banco de dados da AWS
- Amazon DynamoDB Accelerator
- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- 10 melhores dicas de desempenho do Amazon Athena
- Práticas recomendadas do Amazon Redshift Spectrum
- Melhores práticas do Amazon DynamoDB
- Escolha entre o EC2 e o Amazon RDS
- Melhores práticas para a implementação do Amazon ElastiCache

Vídeos relacionados:

- AWS purpose-built databases (DAT209-L) (Bancos de dados com propósito específico da AWS (DAT209-L))
- Amazon Aurora storage demystified: How it all works (DAT309-R) (Armazenamento desmistificado do Amazon Aurora: Como tudo funciona (DAT309-R))
- Amazon DynamoDB deep dive: Advanced design patterns (DAT403-R1)

Exemplos relacionados:

- Optimize Data Pattern using Amazon Redshift Data Sharing (Otimizar padrão de dados usando o compartilhamento de dados do Amazon Redshift)
- Migrações de bancos de dados
- MS SQL Server: Demonstração de replicação do AWS Database Migration Service (DMS)
- Workshop prático de modernização de bancos de dados
- Amostras do Amazon Neptune

PERF04-BP02 Avaliar as opções disponíveis

Compreenda as opções de bancos de dados e como elas podem otimizar a performance antes de você selecionar a sua solução de gerenciamento de dados. Use testes de carga para identificar as métricas de banco de dados que são importantes para a sua workload. Ao explorar as opções de bancos de dados, considere vários aspectos, como grupos de parâmetros, opções de armazenamento, memória, computação, réplica de leitura, consistência eventual, pooling de conexão e opções de armazenamento em cache. Experimente com essas várias opções de configuração para melhorar as métricas.

Resultado desejado: Uma workload pode ter uma ou mais soluções de banco de dados usadas com base nos tipos de dados. A funcionalidade e os benefícios do banco de dados correspondem de maneira ideal as características dos dados, os padrões de acesso e os requisitos da workload. Para otimizar a performance e o custo do banco de dados, você deve avaliar os padrões de acesso aos dados para determinar as opções de banco de dados apropriadas. Avalie os tempos de consulta aceitáveis para garantir que as opções de bancos de dados atendam aos requisitos.

Antipadrões comuns:

- Não identificação dos padrões de acesso aos dados.
- Não ter ciência das opções de configuração da solução de gerenciamento de dados escolhida.

- Contar somente com o aumento do tamanho da instância sem examinar outras opções de configuração.
- Não testar as características de escalabilidade da solução escolhida.

Benefícios do estabelecimento desta prática recomendada: A exploração e a experimentação das opções de bancos de dados permitem que você reduza o custo da infraestrutura, melhore a performance e a escalabilidade e diminua o esforço necessário para manter suas workloads.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

- Precisar otimizar para um banco de dados de tamanho único significa fazer compromissos desnecessários.
- Custos mais altos como resultado da não configuração da solução de banco de dados para que corresponda aos padrões de tráfego.
- Podem surgir problemas operacionais por causa da escalabilidade.
- Os dados podem não estar protegidos no nível necessário.

Orientações para a implementação

Compreenda as características dos dados da sua workload para poder configurar as opções de seu banco de dados. Execute testes de carga para identificar as métricas-chave de performance e os gargalos. Use essas características e métricas para avaliar as opções do banco de dados e experimentar com diferentes configurações.

Serviços da AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
Escalabilidade da computação	Aumente o tamanho das instâncias, as instância	Escalabilidade automática de leitura/gravação com	Aumentar o tamanho das instâncias	Aumentar o tamanho das instâncias, adicionar	Aumentar o tamanho das instâncias	Escala automaticamente para ajustar a	Escalabilidade automática de leitura/gravação com	Escala automaticamente para ajustar a

Serviços da AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
	s do Aurora Serverless escalam automaticamente em resposta a mudanças na carga	modo de capacidade sob demanda ou escalabilidade automática da capacidade de leitura/gravação provisionada em modo de capacidade provisionada		nós ao cluster		capacidade	modo de capacidade sob demanda ou escalabilidade automática da capacidade de leitura/gravação provisionada em modo de capacidade provisionada	capacidade

Serviços da AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
Aumento da escala de leituras horizontalmente	Todos os mecanismos são compatíveis com réplicas de leitura. O Aurora é compatível com a escalabilidade automática de instâncias de réplicas de leitura.	Aumentar as unidades de capacidade de leitura provisionadas	Réplicas de leitura	Réplicas de leitura	Réplicas de leitura. Compatível com a escalabilidade automática de instâncias de réplicas de leitura	Escala automaticamente	Aumentar as unidades de capacidade de leitura provisionadas	Automaticamente aumenta a escala verticalmente para limites de simultaneidade documentados

Serviços da AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
Aumento da escala de gravações horizontalmente	Aumentar o tamanho da instância, realizando gravações em lote na aplicação ou adicionando uma fila na frente do banco de dados. A escalabilidade horizontal via fragmentação em nível de aplicação entre	Aumentar as unidades de capacidade de leitura provisionadas. Garantir a chave de partição ideal para evitar o controle de utilização de gravação em nível de partição	Aumento do tamanho da instância primária	Usar o Redis em modo de cluster para distribuir gravações entre fragmentos	Aumentar o tamanho das instâncias	As solicitações de gravação podem ser limitadas durante a escalabilidade. Se você encontrar exceções de controle de utilização, continue a enviar dados no mesmo throughput (ou mais alto) para escalar automatic	Aumentar as unidades de capacidade de leitura provisionadas. Garantir a chave de partição ideal para evitar o controle de utilização de gravação em nível de partição	Automaticamente aumenta a escala verticalmente para limites de simultaneidade documentados

Serviços da AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
	várias instâncias					amente. Gravaçõe em lote para reduzir solicitações de gravação simultâneas		
Configuração do mecanism	Grupos de parâmetros	Não aplicável	Grupos de parâmetros	Grupos de parâmetros	Grupos de parâmetros	Não aplicável	Não aplicável	Não aplicável

Serviços da AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
Armazenamento em cache	Armazenamento em cache na memória, configurável por meio de grupos de parâmetros. Emparelhar com um cache dedicado, como o ElastiCache for Redis, para descarregar solicitações de itens acessados comumen	Cache totalmente gerenciado do DAX (DAX) disponível	Armazenamento em cache na memória. Opcionalmente, emparelhar com um cache dedicado, como o ElastiCache for Redis, para descarregar solicitações de itens acessados comumen	A função primária é o armazenamento em cache	Usar cache dos resultados de consultas somente leitura para armazenar o resultado em cache	O Timestream tem duas camadas de armazenamento, uma delas é uma camada na memória de alta performance	Implantar um cache dedicado separado, como o ElastiCache for Redis, para descarregar solicitações de itens acessados comumen	Não aplicável

Serviços da AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
Alta disponibilidade/ recuperação de desastres	A configuração recomendada para workloads de produção é a execução de uma instância em espera em uma segunda zona de disponibilidade, para fornecer resiliência em uma região. Para obter resiliência entre regiões, o	Altamente disponível em uma região. As tabelas podem ser replicadas entre regiões usando as tabelas globais do DynamoD	Crie várias instâncias entre zonas de disponibilidade para obter disponibilidade. Os snapshots podem ser compartilhados entre regiões, e os clusters podem ser replicados usando o DMS para fornecer replicaçã	A configuração recomendada para clusters de produção é criar pelo menos um nó em uma zona de disponibilidade secundária. É possível usar o datastore global do ElastiCache para replicar clusters entre regiões.	As réplicas de leitura em outas zonas de disponibilidade funcionam como destinos de failover. Os snapshots podem ser compartilhados entre regiões, e os clusters podem ser replicados usando os	Altamente disponibilidade em uma região. A replicação entre regiões exige o desenvolvimento de uma aplicação personalizada usando o SDK do Timestream	Altamente disponível em uma região. A replicação entre regiões exige a lógica de uma aplicação personalizada ou ferramentas de terceiros	Altamente disponível em uma região. Para replicação entre regiões, exporte o conteúdo do diário do Amazon QLDB para um bucket do S3 e configure o bucket para replicação entre regiões.

Serviços da AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
	Aurora Global Database pode ser usado		o entre regiões/ recuperação de desastres.		streams do Neptune para replicação de dados entre dois clusters em duas regiões diferentes.			

Etapas da implementação

1. Quais opções de configuração estão disponíveis para os bancos de dados selecionados?
    a. Os grupos de parâmetros do Amazon RDS e do Aurora permitem ajustar as configurações em nível de mecanismo comum de bancos de dados, como a memória alocada para o cache ou o ajuste do fuso horário do banco de dados.
    b. Para serviços de bancos de dados provisionados, como o Amazon RDS, o Aurora, o Neptune, o Amazon DocumentDB e os implantados no Amazon EC2, é possível alterar o tipo de instância, o armazenamento provisionado e as réplicas de leitura.
    c. O DynamoDB permite especificar dois modos de capacidade: sob demanda e provisionado. Para levar workloads diferentes em conta, é possível alterar entre esses modos e aumentar a capacidade alocada em modo provisionado a qualquer momento.
2. A leitura ou a gravação da workload é pesada?

a. Quais são as soluções disponíveis para descarregar leituras (réplicas de leitura, armazenamento em cache etc.)?

   i. Em tabelas do DynamoDB, é possível descarregar leituras usando o DAX para armazenamento em cache.

   ii. Em bancos de dados relacionais, é possível criar um cluster do ElastiCache for Redis e configurar a aplicação para ler no cache primeiro, recorrendo ao banco de dados se o item não estiver presente.

   iii. Todos os bancos de dados relacionais, como o Amazon RDS e o Aurora, e os bancos de dados NoSQL provisionados, como o Neptune e o Amazon DocumentDB, são compatíveis com a adição de replicas de leitura para descarregar as partes de leitura da workload.

   iv. Os bancos de dados de tecnologia sem servidor, como o DynamoDB, escalarão automaticamente. Verifique se você tem unidades de capacidade de leitura suficientes (RCU) provisionadas para tratar a workload.

b. Quais são as soluções disponíveis para escalar gravações (fragmentação de chave de partição, introdução de uma fila etc.)?

   i. No caso de bancos de dados relacionais, é possível aumentar o tamanho da instância para acomodar uma workload maior, ou aumentar as IOPs provisionadas para permitir um throughput mais alto para o armazenamento subjacente.

      - Também é possível introduzir uma fila na frente do banco de dados, em vez de gravar diretamente no banco de dados. Esse padrão permite desacoplar a ingestão do banco de dados e controlar a taxa de fluxo, para que o banco de dados não fique sobrecarregado.
      - O uso solicitações de gravação em lote em vez de criar muitas transações de curta duração pode ajudar a melhorar o throughput em bancos de dados relacionais de alto volume de gravação.

   ii. Os bancos de dados de tecnologia sem servidor, como o DynamoDB, podem escalar o throughput de gravação automaticamente ou ajustar as unidades da capacidade de gravação (WCU) provisionadas, dependendo do modo da capacidade.

      - Mas ainda pode ser possível encontrar problemas com partições quentes, ao atingir os limites do throughput de uma determinada chave de partição. Isso pode ser mitigado ao escolher uma chave de partição mais igualmente distribuída ou fragmentar a gravação da chave de partição.

3. Quais são os picos de transações por segundo (TPS) atuais ou esperados? Teste usando esse volume de tráfego e esse volume +X% para compreender as características da escalabilidade.

   a. Ferramentas nativas, como a pg_bench do PostgreSQL, podem ser usadas para testes de estresse do banco de dados e para compreender os gargalos e as características de escalabilidade.
   b. Tráfego do tipo produção deve ser capturado para que possa ser reproduzido para a simulação das condições reais, além de workloads sintéticas.
4. Ao usar computação de tecnologia sem servidor ou escalável elasticamente, teste o impacto da escalabilidade disso no banco de dados. Se apropriado, introduza agrupamento ou gerenciamento de conexões para reduzir o impacto no banco de dados.
   a. O RDS Proxy pode ser usado com o Amazon RDS e o Aurora para gerenciar as conexões ao banco de dados.
   b. Bancos de dados de tecnologia sem servidor, como o DynamoDB, não têm conexões associadas a eles, mas considere a capacidade provisionada e as políticas de escalabilidade automática para lidar com picos na carga.
5. A carga é previsível? Há picos na carga e períodos de inatividade?
   a. Se houver períodos de inatividade, considere reduzir a escala verticalmente da capacidade provisionada ou o tamanho da instância durante esses períodos. Aurora Serverless V2 aumentará e reduzirá a escala verticalmente com base na carga.
   b. No caso de instâncias que não são de produção, considere pausá-las ou interrompê-las durante os horários sem trabalho.
6. É necessário segmentar e separar seus modelos de dados com base nos padrões de acesso e nas características dos dados?
   a. Considere usar o AWS DMS ou o AWS SCT para mover os dados para outros serviços.

Nível de esforço do plano de implementação:

Ao estabelecer essa prática recomendada, lembre-se das características e das métricas atuais dos dados. A coleta dessas métricas, o estabelecimento de uma linha de base e o uso dessas métricas para identificar as opções ideais de configuração de bancos de dados é baixo to moderado . Isso é melhor validade com testes de carga e experimentação.

Recursos

Documentos relacionados:

- Bancos de dados em nuvem com a AWS
- Armazenamento em cache de banco de dados da AWS

- Amazon DynamoDB Accelerator
- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- 10 melhores dicas de desempenho do Amazon Athena
- Práticas recomendadas do Amazon Redshift Spectrum
- Melhores práticas do Amazon DynamoDB

Vídeos relacionados:

- AWS purpose-built databases (DAT209-L) (Bancos de dados com propósito específico da AWS (DAT209-L))
- Amazon Aurora storage demystified: How it all works (DAT309-R) (Armazenamento desmistificado do Amazon Aurora: Como tudo funciona (DAT309-R))
- Amazon DynamoDB deep dive: Advanced design patterns (DAT403-R1)

Exemplos relacionados:

- Exemplos do Amazon DynamoDB
- Amostras de migração de bancos de dados da AWS
- Workshop de modernização de banco de dados
- Como trabalhar com parâmetros no Amazon RDS para bancos de dados Postgress

PERF04-BP03 Coletar e registrar métricas de performance do banco de dados

Para compreender como está a performance dos sistemas de gerenciamento de dados, é importante rastrear métricas relevantes. Essas métricas ajudam a otimizar seus recursos de gerenciamento de dados para garantir que os requisitos da workload sejam atendidos, e que você tenha uma visão geral clara de como está a performance da sua workload. Use ferramentas, bibliotecas e sistemas que registram as medidas de performance relacionadas ao banco de dados.

Há métricas relacionadas ao sistema que hospeda o banco de dados (como, CPU, armazenamento, memória, IOPS), e há métricas para avaliar os próprios dados (como, transações por segundo, taxas de consultas, tempos de resposta, erros). Essas métricas devem estar prontamente acessíveis para

qualquer equipe de suporte e operacional, e devem ter registro histórico suficiente para que seja possível identificar tendências, anomalias e gargalos.

Resultado desejado: Para monitorar a performance das workloads de seus bancos de dados, registre várias métricas de performance ao longo de um período. Isso permite detectar anomalias e avaliar a performance em relação às métricas de negócios para garantir que as necessidades da workload sejam atendidas.

Antipadrões comuns:

- Você só usa a pesquisa manual de arquivos de log para métricas.
- Você só publica métricas para ferramentas internas usadas pela equipe e não tem uma imagem abrangente da workload.
- Você só usa as métricas padrão registradas pelo software de monitoramento selecionado.
- Você só analisa as métricas quando há um problema.
- Você só monitora as métricas em nível do sistema, não captura as métricas de acesso aos dados e de uso.

Benefícios do estabelecimento desta prática recomendada: O estabelecimento de uma linha de base de performance ajuda a compreender o comportamento normal e os requisitos das workloads. Padrões anormais podem ser identificados e depurados mais rapidamente, melhorando a performance e a confiabilidade do banco de dados. A capacidade do banco de dados pode ser configurada para garantir um custo ideal sem comprometer a performance.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

- A inabilidade de diferenciar o nível de performance fora do normal versus normal criam dificuldades na identificação de problemas e na tomada de decisões.
- A economia de custo possível pode não ser identificada.
- Os padrões de crescimento não serão identificados resultando em degradação da confiabilidade e da performance.

Orientações para a implementação

Identifique, colete, agregue e correlacione métricas relacionadas ao banco de dados. As métricas devem incluir as métricas do sistema subjacente que oferece suporte ao banco de dados e do banco

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

Etapas da implementação:

1. Quais métricas de bancos de dados é importante rastrear?
    a. Métricas de monitoramento do Amazon RDS
    b. Monitoramento com insights da performance
    c. Monitoramento aprimorado
    d. Métricas do DynamoDB
    e. Monitoramento do DynamoDB DAX
    f. Monitoramento do MemoryDB
    g. Monitoramento do Amazon Redshift
    h. Métricas e dimensões de séries temporais
    i. Métricas em nível de cluster do Aurora
    j. Monitoramento do Amazon Keyspaces
    k. Monitoramento do Amazon Neptune
2. O monitoramento do banco de dados se beneficiaria de uma solução de machine learning que detecta problemas de performance por anomalias operacionais?
    a. O Amazon DevOps Guru for Amazon RDS fornece visibilidade dos problemas de performance e faz recomendações de ações corretivas.
3. São necessários detalhes em nível de aplicação sobre o uso do SQL?
    a. AWS X-Ray pode ser instrumentado na aplicação para obter insights e encapsular todos os pontos de dados em uma única consulta;
4. Você tem uma solução de registro em log e de monitoramento aprovada?
    a. Amazon CloudWatch pode coletar métricas nos recursos na sua arquitetura. Você também pode coletar e publicar métricas personalizadas para descobrir métricas de negócio ou derivadas. Use o CloudWatch ou soluções de terceiros para definir alarmes que indiquem quando os limites são violados.

5. Você identificou e configurou suas políticas de retenção de dados para que correspondam às suas metas operacionais e de segurança?
   a. Retenção de dados padrão para métricas do CloudWatch
   b. Retenção de dados padrão do CloudWatch Logs

Nível de esforço do plano de implementação: Há um nível de esforço médio para identificar, rastrear, coletar, agregar e correlacionar métricas de todos os recursos de banco de dados.

Recursos

Documentos relacionados:

- Armazenamento em cache de banco de dados da AWS
- 10 melhores dicas de desempenho do Amazon Athena
- Práticas recomendadas do Amazon Aurora
- Amazon DynamoDB Accelerator
- Melhores práticas do Amazon DynamoDB
- Práticas recomendadas do Amazon Redshift Spectrum
- Performance do Amazon Redshift
- Bancos de dados em nuvem com a AWS
- Insights de performance do Amazon RDS

Vídeos relacionados:

- AWS purpose-built databases (DAT209-L) (Bancos de dados com propósito específico da AWS (DAT209-L))
- Amazon Aurora storage demystified: How it all works (DAT309-R) (Armazenamento desmistificado do Amazon Aurora: Como tudo funciona (DAT309-R))
- Amazon DynamoDB deep dive: Advanced design patterns (DAT403-R1)

Exemplos relacionados:

- Level 100: Monitoring with CloudWatch Dashboards (Nível 100: monitoramento com os painéis do CloudWatch)

- AWS Dataset Ingestion Metrics Collection Framework (Framework da coleção de métricas de ingestão de conjuntos de dados da AWS)
- Amazon RDS Monitoring Workshop (Workshop sobe o monitoramento do AWS RDS)

PERF04-BP04 Escolher armazenamento de dados com base nos padrões de acesso

Use os padrões de acesso da carga de trabalho para decidir que serviços e tecnologias usar. Além dos requisitos não funcionais, como performance e escala, os padrões de acesso influenciam fortemente a escolha das soluções de banco de dados e de armazenamento. A primeira dimensão é a necessidade de transações, conformidade com ACID e leituras consistentes. Nem todo banco de dados é compatível com isso, e a maioria dos bancos de dados NoSQL fornecem um modelo de consistência eventual. A segunda importante dimensão seria a distribuição de gravações e leituras ao longo do tempo e do espaço. As aplicações distribuídas globalmente precisam considerar os requisitos de padrões de tráfego, de latência e de acesso para identificar a solução ideal de armazenamento. O terceiro aspecto essencial a considerar é a flexibilidade dos padrões de consultas, os padrões de acesso aleatório e as consultas de uma única vez. Considerações relativas à funcionalidade de consultas altamente especializadas para processamento de texto e de linguagem natural, às séries temporais e aos grafos também devem ser avaliadas.

Resultado desejado: O armazenamento de dados foi selecionado com base nos padrões de acesso aos dados identificados e documentados. Isso pode incluir as consulta mais comuns de leitura, gravação e exclusão, as necessidades de cálculos e agregações ad-hoc, a complexidade dos dados, a interdependência dos dados e a consistência exigida.

Antipadrões comuns:

- Você só seleciona um fornecedor de banco de dados para simplificar o gerenciamento de operações.
- Você pressupõe que os padrões de acesso aos dados permanecerão consistentes ao longo do tempo.
- Você implementa transações complexas, reversão e lógica de consistência na aplicação.
- O banco de dados está configurado para ser compatível com alta expansão potencial de tráfego, o que faz com que os recursos do banco de dados fiquem ociosos a maior parte do tempo.
- O uso de um banco de dados compartilhado para usos transacionais e analíticos.

Benefícios do estabelecimento desta prática recomendada: A seleção e a otimização do armazenamento de dados com base em padrões de acesso ajuda a reduzir a complexidade do

desenvolvimento e a otimizar as oportunidades de performance. A compreensão de quando usar réplicas de leitura, tabelas globais e armazenamento em cache ajuda a reduzir a sobrecarga operacional e a escalar com base nas necessidades da workload.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

Identifique e avalie o padrão de acesso aos dados para selecionar a configuração correta do armazenamento. Cada solução de banco de dados tem opções para configurar e otimizar sua solução de armazenamento. Use as métricas e os logs coletados e experimente com opções para encontrar a configuração ideal. Use a tabela a seguir para analisar as opções de armazenamento por serviço de banco de dados.

Serviços da AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
Escalabilidade de armazenamento.	A opção de escalabilidade automática de armazenamento disponível para escalar automaticamente IOPS provisionadas de armazenamento também	Escala automaticamente. As tabelas não são restringidas em termos de tamanho.	A opção de escalabilidade automática de armazenamento disponível escala o armazenamento provisionado.	O armazenamento é na memória, vinculado ao tipo ou à contagem de instâncias.	A opção de escalabilidade automática de armazenamento disponível para escalar automaticamente o armazenamento provisionado.	Configura o período de retenção para camadas na memória e magnética s em dias.	Aumenta verticalmente e reduz horizontalmente a escala do armazenamento de tabelas automaticamente.	Escala automaticamente. As tabelas não são restringidas em termos de tamanho.

Serviços da AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
	pode ser usada independentemente do armazenamento provisionado ao utilizar tipos de IOPs provisionadas de armazenamento.							

Etapas da implementação:

1. Identifique e documente o crescimento antecipado dos dados e do tráfego.
    a. O Amazon RDS e o Aurora são compatíveis com o aumento da escala vertical automática até os limites documentados. Além disso, considere fazer a transição de dados mais antigos para arquivamento do Amazon S3, agregando dados históricos para análise ou escalando horizontalmente por meio de fragmentação.
    b. O DynamoDB e o Amazon S3 escalam até o volume de armazenamento quase ilimitado automaticamente.
    c. As instâncias e os bancos de dados do Amazon RDS em execução no EC2 podem ser redimensionados manualmente, e as instâncias do EC2 podem ter novos volumes do EBS adicionados posteriormente para armazenamento adicional.

d. Os tipos de instância podem ser alterados com base nas alterações nas atividades. Por exemplo, é possível iniciar com uma instância menor para teste e escalar a instância quando o tráfego de produção começar a ser recebido no serviço. O Aurora Serverless V2 reduz a escala horizontalmente automaticamente em resposta a alterações na carga.

1. Documente os requisitos de performance em condições normais e em picos (transações por segundo (TPS) e consultas por segundo (QPS)) e de consistência (consistência ACID e eventual).
2. Documente os aspectos da implantação da solução e os requisitos de acesso ao banco de dados (global, Multi-AZ, replicação de leitura, vários nós de gravação).

Nível de esforço para o plano de implementação: Se você não tiver logs ou métricas para a solução de gerenciamento dos dados, será necessário realizar isso para identificar e documentar os padrões de acesso aos seus dados. Depois que o padrão de acesso aos dados for compreendido, o nível de esforço para selecionar e configurar o armazenamento dos seus dados é baixo .

Recursos

Documentos relacionados:

- Armazenamento em cache de banco de dados da AWS
- 10 melhores dicas de desempenho do Amazon Athena
- Práticas recomendadas do Amazon Aurora
- Amazon DynamoDB Accelerator
- Melhores práticas do Amazon DynamoDB
- Práticas recomendadas do Amazon Redshift Spectrum
- Performance do Amazon Redshift
- Bancos de dados em nuvem com a AWS
- Tipos de armazenamento do Amazon RDS

Vídeos relacionados:

- AWS purpose-built databases (DAT209-L) (Bancos de dados com propósito específico da AWS (DAT209-L))
- Amazon Aurora storage demystified: How it all works (DAT309-R) (Armazenamento desmistificado do Amazon Aurora: Como tudo funciona (DAT309-R))

- Amazon DynamoDB deep dive: Advanced design patterns (DAT403-R1)

Exemplos relacionados:

- Experimentar e testar com testes de carga distribuída na AWS

PERF04-BP05 Otimizar o armazenamento de dados com base nas métricas e nos 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.

Antipadrões comuns:

- Você só usa a pesquisa manual de arquivos de log para métricas.
- Você só publica métricas em ferramentas internas.

Benefícios do estabelecimento desta prática recomendada: Para garantir que você esteja atendendo às métricas necessárias para a carga de trabalho, você deve monitorar as métricas de performance do banco de dados relacionadas a leituras e gravações. Você pode usar esses dados para adicionar novas otimizações para leituras e gravações na camada de armazenamento de dados.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

Otimizar o armazenamento de dados com base em métricas e padrões: use as métricas relatadas para identificar todas as áreas com baixa performance na sua workload e otimizar os componentes do banco de dados. Cada sistema de banco de dados tem características relacionadas a performance diferentes a serem avaliadas, como a maneira de indexar, armazenar em cache ou distribuir os dados entre vários sistemas. Meça o impacto de suas otimizações.

Recursos

Documentos relacionados:

- Armazenamento em cache de banco de dados da AWS

- 10 melhores dicas de desempenho do Amazon Athena
- Práticas recomendadas do Amazon Aurora
- Amazon DynamoDB Accelerator
- Melhores práticas do Amazon DynamoDB
- Práticas recomendadas do Amazon Redshift Spectrum
- Performance do Amazon Redshift
- Bancos de dados em nuvem com a AWS
- Analisar as anomalias da performance com o DevOps Guru for RDS
- Modo de capacidade de leitura/gravação do DynamoDB

Vídeos relacionados:

- AWS purpose-built databases (DAT209-L) (Bancos de dados com propósito específico da AWS (DAT209-L))
- Amazon Aurora storage demystified: How it all works (DAT309-R) (Armazenamento desmistificado do Amazon Aurora: Como tudo funciona (DAT309-R))
- Amazon DynamoDB deep dive: Advanced design patterns (DAT403-R1)

Exemplos relacionados:

- Laboratórios práticos do Amazon DynamoDB

PERF 5 Como você configura sua solução de rede?

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.

Práticas recomendadas

- PERF05-BP01 Compreender como as redes afetam a performance
- PERF05-BP02 Avaliar os recursos de redes disponíveis
- PERF05-BP03 Escolher a VPN ou a conectividade dedicada dimensionada adequadamente para workloads híbridas

- PERF05-BP04 Utilizar o balanceamento de carga e o descarregamento da criptografia
- PERF05-BP05 Escolher os protocolos de rede para aumentar a performance
- PERF05-BP06 Escolher o local da sua workload com base nos requisitos de rede
- PERF05-BP07 Otimizar a configuração da rede com base em métricas

PERF05-BP01 Compreender como as redes afetam a performance

Analise e entenda como decisões relacionadas à rede afetam a performance da carga de trabalho. A rede é responsável pela conectividade entre os componentes da aplicação, os serviços de nuvem, as redes de borda e os dados on-premises, e, portanto, ela pode afetar significativamente a performance da workload. Além da performance da workload, a experiência dos usuários também é afetada pela latência da rede, a largura de banda, os protocolos, a localização, a congestão da rede, a tremulação, o throughput e as regras de roteamento.

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

Antipadrões comuns:

- Todo o tráfego flui por meio dos datacenters existentes.
- Você cria sessões do Direct Connect em excesso sem compreender os requisitos reais de uso.
- Você não considera as características da workload e a sobrecarga da criptografia ao definir suas soluções de redes.
- Você usa conceitos e estratégias de on-premises para soluções de redes na nuvem.

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

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

Identifique as métricas importantes de performance da rede da sua workload e capture as características da rede. Defina e documente os requisitos como parte de uma abordagem direcionada por dados, usando testes comparativos ou de carga. Use esses dados para identificar onde sua solução de rede é restringida e examine as opções de configuração que podem melhorar a workload. Compreenda as opções e os recursos de redes nativos da nuvem disponíveis e como eles podem afetar a performance da workload com base nos requisitos. Cada recurso de redes tem vantagens e desvantagens, e pode ser configurado para atender às características da sua workload e escalar com base em suas necessidades.

Etapas da implementação:

1. Defina e documente os requisitos de performance das redes:
    a. Inclua métricas, como as de latência de rede, largura de banda, protocolos, locais, padrões de tráfego (picos e frequências), throughput, criptografia, inspeção e regras de roteamento
2. Capture as características fundamentais das redes:
    a. Logs de fluxo da VPC
    b. Métricas do AWS Transit Gateway
    c. Métricas do AWS PrivateLink
3. Capture as características das redes da sua aplicação:
    a. Adaptador de rede elástica
    b. Métricas do AWS App Mesh
    c. Métricas do Amazon API Gateway
4. Capture as características das redes de borda:
    a. Métricas do Amazon CloudFront
    b. Métricas do Amazon Route 53
    c. Métricas do AWS Global Accelerator
5. Capture as características das redes híbridas:
    a. Métricas do Direct Connect
    b. Métricas do AWS Site-to-Site VPN
    c. Métricas da AWS Client VPN
    d. Métricas da WAN da Nuvem AWS
6. Capture as características das redes de segurança:

   a. Métricas do AWS Shield, do WAF e do Network Firewall

7. Capture as métricas de performance de ponta a ponta com ferramentas de rastreamento:
   a. AWS X-Ray
   b. Amazon CloudWatch RUM
8. Realize teste comparativo e teste de performance da rede:
   a. Realize o teste comparativo do throughput da rede: alguns fatores que podem afetar a performance da rede do EC2 quando as instâncias estão na mesma VPC; Meça a largura de banda da rede entre as instâncias do EC2 do Linux na mesma VPC.
   b. Execute testes de carga para experimentar com soluções e opções de redes

Nível de esforço do plano de implementação: Há um nível de esforço médio para documentar os requisitos, as opções e as soluções disponíveis de redes da workload.

Recursos

Documentos relacionados:

- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Elastic Network Adapter (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- Transit Gateway
- Fazer a transição para o roteamento baseado em latência no Amazon Route 53
- VPC Endpoints
- Logs de fluxo da VPC

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1) (Conectividade com a AWS e arquiteturas de rede híbrida da AWS (NET317-R1)

- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)
- Melhorar a performance da rede global para aplicações
- EC2 Instances and Performance Optimization Best Practices (Práticas recomendadas para instâncias do EC2 e otimização da performance)
- Optimizing Network Performance for Amazon EC2 Instances (Otimização da performance da rede para instâncias do Amazon EC2)
- Networking best practices and tips with the Well-Architected Framework (Práticas recomendadas e dicas de redes com o Well-Architected Framework)
- AWS networking best practices in large-scale migrations (Práticas recomendadas da AWS em migrações de grande escala)

Exemplos relacionados:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluções de segurança escaláveis)
- AWS Networking Workshops

PERF05-BP02 Avaliar os recursos de redes 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 perda de pacotes ou a instabilidade da rede.

Vários serviços são criados para melhorar a performance e outros costumam oferecer recursos para otimizar a performance da rede. Serviços como o AWS Global Accelerator e Amazon CloudFront existem para melhorar a performance enquanto a maioria dos outros serviços têm recursos para otimizar o tráfego de rede. Analise os recursos do serviço, como a capacidade da rede da instância do EC2, os tipos de instância para redes aprimoradas, as instâncias otimizadas para Amazon EBS, a aceleração de transferências do Amazon S3 e o CloudFront, para melhorar a performance da workload.

Resultado desejado: você documentou o inventário de componentes da workload e identificou quais configurações de rede vão ajudar cada componente com o intuito de atender aos seus requisitos de performance. Depois de avaliar os recursos de rede, você testou e mensurou as métricas de performance para identificar como usar os recursos disponíveis.

Antipadrões comuns:

- Você coloca todas as workloads na Região da AWS que está mais próxima de sua sede em vez de uma Região da AWS próxima dos usuários finais.
- Não obter uma referência para a performance de sua workload e deixar de avaliar continuamente a performance de sua workload em relação a essa referência.
- Você não avalia configurações de serviço para opções de melhoria da performance.

Benefícios do estabelecimento desta prática recomendada: Avaliar todos os recursos e opções de serviços pode aumentar a performance de sua workload, reduzir o custo da infraestrutura, diminuir o esforço necessário para manter sua workload e aumentar sua postura geral de segurança. É possível utilizar a espinha dorsal da AWS para garantir a experiência ideal de redes para os clientes.

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

Orientação de implementação

Analise quais opções de configuração de rede estão disponíveis e como elas poderiam afetar sua workload. Entender como essas opções interagem com sua arquitetura e o impacto que elas terão sobre a performance medida e a performance percebida pelos usuários é fundamental para a otimização da performance.

Etapas da implementação:

1. Crie uma lista de componentes da workload.
   a. Desenvolva, gerencie e monitore a rede de sua organização usando o Nuvem AWS WAN.
   b. Obtenha visibilidade de sua rede usando o Network Manager. Use uma ferramenta de banco de dados de gerenciamento de configurações (CMDB) existente ou uma ferramenta como o AWS Config para criar um inventário de sua workload e como ela é configurada.
2. Se for uma workload existente, identifique e documente a referência para suas métricas de performance, focando nos gargalos e nas áreas de melhoria. As métricas de rede associadas a performance vão variar de acordo com a workload com base nos requisitos comerciais e nas características da workload. Como ponto de partida, a análise dessas métricas pode ser importante para sua workload: largura de banda, latência, perda de pacotes, instabilidade da rede e retransmissões.
3. Se a workload for nova, realize testes de carga para identificar gargalos de performance.
4. Para os gargalos de performance que identificar, analise as opções de configuração para suas soluções a fim de identificar oportunidades de melhoria da performance.

5. Se não souber seu caminho ou rotas de rede, use o Network Access Analyzer para identificá-los.
6. Revise seus protocolos de rede para reduzir ainda mais a latência.
   - PERF05-BP05 Escolher os protocolos de rede para aumentar a performance
7. Se você usar uma AWS Site-to-Site VPN em vários locais para se conectar a uma Região da AWS, analise as conexões aceleradas de Site-to-Site VPN em busca de oportunidades para melhorar a performance de rede.
8. Quando o tráfego da workload está distribuído entre várias contas, avalie a topologia de rede e os serviços para reduzir a latência.
   - Avalie suas concessões de operação e performance entre Emparelhamento de VPC e AWS Transit Gateway ao conectar várias contas. O AWS Transit Gateway oferece suporte a uma throughput de AWS Site-to-Site VPN para dimensionamento além de um único limite máximo de IPsec usando vários caminhos. O tráfego entre uma Amazon VPC e o AWS Transit Gateway permanece na rede privada da AWS e não é exposto à Internet. O AWS Transit Gateway simplifica a interconexão de todas as suas VPCs, o que pode abranger milhares de Contas da AWS e redes on-premises. Compartilhe seu AWS Transit Gateway entre várias contas usando o Resource Access Manager. Para obter visibilidade de seu tráfego de rede global, use o Network Manager para obter uma visão central de suas métricas de rede.
9. Revise seus locais de usuários e minimize a distância entre os usuários e a workload.
   a. AWS Global Accelerator é um serviço de rede que melhora a performance do tráfego dos usuários em até 60% usando a infraestrutura de rede global da Amazon Web Services. Quando a Internet está congestionada, o AWS Global Accelerator otimiza o caminho para a aplicação a fim de manter a perda de pacotes, instabilidade da rede e latência num nível consistentemente baixo. Também fornece endereços IP estáticos que simplificam a movimentação de endpoints entre zonas de disponibilidade ou Regiões da AWS sem necessidade de atualizar sua configuração de DNS ou alterar as aplicações voltadas para os clientes.
   b. Amazon CloudFront pode melhorar a performance de entrega de conteúdo de sua workload e latência globalmente. O CloudFront tem mais de 410 pontos de presença distribuídos globalmente para armazenar seu conteúdo em cache e reduzir a latência para o usuário final.
   c. O Amazon Route 53 oferece opções de roteamento baseado em latência, roteamento por geolocalização, roteamento por geoproximidadee aos roteamento baseado em IP para ajudar a melhorar a performance de sua workload para um público global. Identifique qual opção de roteamento otimiza a performance da workload analisando o tráfego e a localização dos usuários da workload.
10. Avalie recursos adicionais do Amazon S3 para melhorar as IOPS de armazenamento.

a. Aceleração de Transferências do Amazon S3 é um recurso que permite que usuários externos se beneficiem de otimizações de rede do CloudFront a fim de fazer upload de dados no Amazon S3. Isso melhora a capacidade de transferir grandes quantidades de dados com origem em locais remotos que não têm conectividade dedicada com a Nuvem AWS.

b. Pontos de acesso multirregionais no Amazon S3 replicam conteúdo para várias regiões e simplificam a workload ao proporcionar um ponto de acesso. Quando um ponto de acesso multirregional é usado, você pode solicitar ou gravar dados no Amazon S3 com o serviço identificando o bucket de menor latência.

11 Revise a largura de banda da rede para recursos de computação.

a. Interfaces de rede elástica (ENI) usadas por instâncias do EC2, contêineres e funções do Lambda são limitadas por fluxo. Revise seus grupos de posicionamento para otimizar a throughput de rede do EC2. Para evitar gargalos na abordagem por fluxo, projete sua aplicação para usar vários fluxos. Para monitorar e obter visibilidade de suas métricas de rede relacionadas à computação, use métricas do CloudWatch e `ethtool`. `ethtool` está incluído no driver de ENI e expõe métricas adicionais relacionadas à rede que podem ser publicadas como métricas personalizadas no CloudWatch.

b. As instâncias mais novas do EC2 também podem aproveitar as redes aprimoradas. Instâncias série N do EC2, como `M5n` e `M5dn`, utilizam a quarta geração de cartões Nitro personalizados para oferecer até 100 Gbps de throughput de rede a uma única instância. Essas instâncias oferecem quatro vezes mais largura de banda de rede e processo de pacotes em comparação às instâncias `M5` básicas, sendo ideais para aplicações com uso intenso de rede.

c. Adaptadores de rede elástica (ENA) da Amazon proporcionam ainda mais otimização ao oferecer mais throughput para suas instâncias em um grupo de posicionamento de cluster.

d. Elastic Fabric Adapter é uma interface de rede para instâncias do Amazon EC2 que permite executar workloads que exigem altos níveis de comunicação entre nós em grande escala na AWS. Com o EFA, os aplicativos de Computação de Alta Performance (HPC) que usam a Message Passing Interface (MPI – Interface de Passagem de Mensagens) e os aplicativos de ML (Machine Learning) que usam a NCCL (NVIDIA Collective Communications Library) podem aumentar a escala para milhares de CPUs ou GPUs.

e. Instâncias otimizadas para Amazon EBS usam uma pilha de configuração otimizada e fornecem capacidade adicional e dedicada para E/S do Amazon EBS. Essa otimização proporciona a melhor performance para seus volumes do EBS ao minimizar a contenção entre a E/S do Amazon EBS e outros tráfegos provenientes da sua instância.

Nível de esforço do plano de implementação:

Para estabelecer esta prática recomendada, você deve conhecer as opções atuais de componentes para sua workload que afetam a performance de rede. Reunir os componentes, avaliar as opções de melhoria da rede, realizar testes, implementar e documentar essas melhorias são ações de esforço baixo to moderado .

Recursos

Documentos relacionados:

- Instâncias otimizadas para Amazon EBS
- Application Load Balancer
- Largura de banda da rede para instâncias do Amazon EC2
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Elastic Network Adapter (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- AWS Transit Gateway
- Fazer a transição para o roteamento baseado em latência no Amazon Route 53
- VPC Endpoints
- Logs de fluxo da VPC
- Como criar um CMDB na nuvem
- Dimensionar a throughput de VPN usando o AWS Transit Gateway

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1) (Conectividade com a AWS e arquiteturas de rede híbrida da AWS)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)
- AWS Global Accelerator

Exemplos relacionados:

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

PERF05-BP03 Escolher a VPN ou a conectividade dedicada dimensionada adequadamente para workloads híbridas

Quando uma rede comum é necessária para conectar recursos on-premises e de nuvem na AWS, verifique se você tem largura de banda adequada para atender aos requisitos de performance. Estime os requisitos de largura de banda e de latência para a sua workload híbrida. Esses números orientarão os requisitos de dimensionamento para o AWS Direct Connect ou seus endpoints de VPN.

Resultado desejado: Ao implantar uma workload que precisa de conectividade de rede híbrida, há várias opções de configuração de conectividade, como o Direct Connect ou as VPNs gerenciadas ou não gerenciadas. Selecione o tipo de conexão apropriado para cada workload e garanta que você tem os requisitos de largura de banda e de criptografia adequados entre seu local e a nuvem.

Antipadrões comuns:

- Você só avalia as soluções de VPN para seus requisitos de criptografia de rede.
- Você não avalia opções de conectividade paralela ou de backup.
- Você usa configurações padrão para roteadores, túneis e sessões de BGP.
- Você não compreende ou identifica todos os requisitos da workload (necessidades de criptografia, protocolo, largura de banda e tráfego).

Benefícios do estabelecimento desta prática recomendada: A seleção e a configuração adequadas de soluções de rede híbrida aumentará a confiabilidade da sua workload e maximizará as oportunidades de performance. A identificação dos requisitos da workload, o planejamento antecipado e a avaliação das soluções híbridas ajudarão você a minimizar alterações dispendiosas da rede física e a sobrecarga operacional e reduzirá o tempo de colocação no mercado.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

Desenvolva uma arquitetura de redes híbridas com base nos seus requisitos de largura de banda. Calcule os requisitos de largura de banda e de latência de suas aplicações híbridas. Com base nos requisitos de largura de banda, uma única conexão VPN ou Direct Connect pode não ser suficiente,

e você deve projetar uma configuração híbrida para habilitar o balanceamento de carga de tráfego em várias conexões. O Direct Connect pode ser necessário por oferecer performance mais previsível e consistente devido à conectividade de rede privada dele. Ele é ótimo para cargas de trabalho de produção que exigem latência consistente e quase zero tremulação.

O AWS Direct Connect oferece conectividade dedicada ao ambiente da AWS, de 50 Mbps até 10 Gbps. Isso permite que você tenha latência gerenciada e controlada, além de largura de banda provisionada para que sua carga de trabalho possa se conectar facilmente e com alta performance a outros ambientes. Usando um dos parceiros do AWS Direct Connect, é possível ter conectividade completa de vários ambientes, fornecendo uma rede estendida com performance consistente.

O AWS Site-to-Site VPN é um serviço de VPN gerenciada para VPCs. Quando uma conexão VPN é criada, a AWS fornece túneis para dois endpoints de VPN diferentes. Com o AWS Transit Gateway, você pode simplificar a conectividade entre várias VPCs e também conectar-se a qualquer VPC anexada ao AWS Transit Gateway com uma única conexão VPN. O AWS Transit Gateway também permite escalar além do limite de throughput de IPsec de 1,25 Gbps da VPN ao habilitar o suporte para roteamento equal cost multi-path (ECMP – Caminho múltiplo de custo igual) por vários túneis de VPN.

Nível de esforço para o plano de implementação: Há um nível de esforço alto para avaliar as necessidades da workload para redes híbridas e para implementar soluções de redes híbridas.

Recursos

Documentos relacionados:

- Network Load Balancer
- Produtos de redes com a AWS
- Transit Gateway
- Fazer a transição para o roteamento baseado em latência no Amazon Route 53
- VPC Endpoints
- Logs de fluxo da VPC
- Site-to-Site VPN
- Building a Scalable and Secure Multi-VPC AWS Network Infrastructure (Criação de uma infraestrutura de rede da AWS de várias VPCs escaláveis e seguras)
- Direct Connect

- Client VPN

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1) (Conectividade com a AWS e arquiteturas de rede híbrida da AWS (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)
- AWS Global Accelerator
- Direct Connect
- Conexão do Transit Gateway
- Soluções de VPN
- Segurança com as soluções de VPN

Exemplos relacionados:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluções de segurança escaláveis)
- AWS Networking Workshops

PERF05-BP04 Utilizar o balanceamento de carga e o descarregamento da 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.

Ao implementar uma arquitetura de aumento da escala verticalmente na qual você quer usar várias instâncias para o conteúdo do serviço, use balanceadores de carga na Amazon VPC. A AWS fornece vários modelos para suas aplicações no serviço ELB. O Application Load Balancer é mais adequado para balanceamento de carga de tráfego HTTP e HTTPS e fornece roteamento avançado de solicitações direcionadas na entrega de arquiteturas de aplicações modernas, incluindo microsserviços e contêineres.

O Network Load Balancer é mais adequado para o balanceamento de carga de tráfego TCP em que é necessária performance extrema. Ele é capaz de processar milhões de solicitações por segundo

enquanto mantém latências ultrabaixas, e também é otimizado para lidar com padrões de tráfego súbitos e voláteis.

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

Antipadrões comuns:

- Você roteia todo o tráfego da Internet por meio de load balancers existentes.
- Você usa o balanceamento de carga TCP genérico e faz com que cada nó de computação lide com a criptografia SSL.

Benefícios do estabelecimento desta prática recomendada: Um load balancer lida com a carga variável do tráfego do aplicativo em uma única zona de disponibilidade ou em várias zonas de disponibilidade. Os load balancers oferecem alta disponibilidade, escalabilidade automática e segurança robusta, necessárias para tornar seus aplicativos tolerantes a falhas.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

Usar o balanceador de carga adequado para a sua workload: selecione o balanceador de carga adequado para a sua workload. Se você precisar balancear a carga de solicitações HTTP, recomendamos o Application Load Balancer. Para balanceamento de carga de protocolos da rede e de transporte (camada 4, TCP, UDP) e para aplicações de performance extrema e baixa latência, recomendamos o Network Load Balancer. Os Application Load Balancers são compatíveis com balanceadores de carga HTTPS, e os Network Load Balancers são compatíveis com o descarregamento de criptografia TLS.

Ativar o descarregamento de criptografia HTTPS ou TLS: o Elastic Load Balancing inclui o gerenciamento integrado de certificados, a autenticação de usuários e a descriptografia SSL/TLS. Ele fornece a flexibilidade de gerenciar centralmente as configurações de TLS e descarregar cargas de trabalho com uso intensivo de CPU de seus aplicativos. Criptografe todo o tráfego HTTPS como parte da implantação do load balancer.

Recursos

Documentos relacionados:

- Instâncias otimizadas para Amazon EBS
- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Elastic Network Adapter (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- Transit Gateway
- Fazer a transição para o roteamento baseado em latência no Amazon Route 53
- VPC Endpoints
- Logs de fluxo da VPC

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1) (Conectividade com a AWS e arquiteturas de rede híbrida da AWS (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)

Exemplos relacionados:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluções de segurança escaláveis)
- AWS Networking Workshops

PERF05-BP05 Escolher 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.

Há uma relação entre latência e largura de banda para alcançar o throughput. Se a transferência de arquivos estiver usando TCP, latências mais altas reduzirão o throughput geral. Existem abordagens para corrigir isso com ajuste de TCP e protocolos de transferência otimizados. Algumas abordagens usam UDP.

Antipadrões comuns:

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

Benefícios do estabelecimento desta prática recomendada: Selecionar o protocolo apropriado para comunicação entre os componentes da carga de trabalho garante que você esteja obtendo a melhor performance para essa carga de trabalho. O UDP sem conexão permite alta velocidade, mas não oferece retransmissão ou alta confiabilidade. TCP é um protocolo completo, mas requer maior sobrecarga para processar os pacotes.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

Otimizar o tráfego da rede: selecione o protocolo apropriado para otimizar a performance da sua workload. Há uma relação entre latência e largura de banda para alcançar o throughput. Se a transferência de arquivos estiver usando TCP, latências mais altas reduzirão o throughput geral. Existem abordagens para corrigir a latência com ajuste de TCP e protocolos de transferência otimizados. Algumas abordagens usam UDP.

Recursos

Documentos relacionados:

- Instâncias otimizadas para Amazon EBS
- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Elastic Network Adapter (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- Transit Gateway
- Fazer a transição para o roteamento baseado em latência no Amazon Route 53
- VPC Endpoints
- Logs de fluxo da VPC

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1) (Conectividade com a AWS e arquiteturas de rede híbrida da AWS (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)

Exemplos relacionados:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluções de segurança escaláveis)
- AWS Networking Workshops

PERF05-BP06 Escolher o local da sua workload com base nos requisitos de rede

Use as opções de localização de nuvem disponíveis para reduzir a latência de rede ou aprimorar o throughput. Use as Regiões da AWS, as zonas de disponibilidade, os grupos de posicionamento e os locais de borda, como o AWS Outposts, as zonas locais da AWS e o AWS Wavelength, para reduzir a latência da rede ou melhorar o throughput.

A infraestrutura da Nuvem AWS é criada em torno de regiões e de zonas de disponibilidade. Uma região é um local físico no mundo com várias zonas de disponibilidade.

As zonas de disponibilidade consistem em um ou mais datacenters discretos que estão alojados em instalações separadas, cada uma com energia, rede e conectividade redundantes. Essas zonas de disponibilidade oferecem a capacidade de operar aplicativos e bancos de dados de produção com níveis de disponibilidade, tolerância a falhas e escalabilidade mais altos do que seria possível em um único datacenter

Escolha uma ou mais regiões adequadas para sua implantação conforme nos seguintes elementos fundamentais:

- A localização dos seus usuários: a escolha de uma região próxima de sua carga de trabalho garante menor latência quando eles usarem a carga de trabalho.
- A localização dos seus dados: para aplicativos com uso intenso de dados, o maior gargalo na latência é a transferência de dados. O código do aplicativo deve ser executado o mais perto possível dos dados.
- Outras restrições: leve em conta restrições como segurança e conformidade.

O Amazon EC2 oferece grupos de posicionamento para redes. Um grupo de posicionamento é um agrupamento lógico de instâncias para reduzir a latência ou para aumentar a confiabilidade. O uso de placement groups com tipos de instância compatíveis e um Elastic Network Adapter (ENA) permite que as cargas de trabalho participem de uma rede de 25 Gbps e baixa latência. Recomenda-se o uso de placement groups para cargas de trabalho que se beneficiam de baixa latência de rede, alto throughput de rede ou ambos. O uso de placement groups tem o benefício de reduzir a instabilidade em comunicações de rede.

Serviços sensíveis à latência são entregues na borda usando uma rede global de pontos de presença. Esses pontos de presença costumam oferecer serviços como Content Delivery Network (CDN – Rede de entrega de conteúdo) e Domain Name System (DNS). Ao ter esses serviços na borda, as cargas de trabalho podem responder com baixa latência a solicitações de conteúdo ou resolução de DNS. Esses serviços também fornecem serviços geográficos, como direcionamento geográfico de conteúdo (fornecendo conteúdo diferente conforme o local do usuário final) ou roteamento com base em latência para direcionar os usuários finais à região mais próxima (latência mínima).

Amazon CloudFront é uma CDN global que pode ser usada para acelerar conteúdo estático, como imagens, scripts e vídeos, bem como conteúdo dinâmico, como APIs ou aplicativos web. Ele conta com uma rede global de pontos de presença que armazenarão o conteúdo em cache e fornecerão conectividade de rede de alta performance aos seus usuários. O CloudFront também acelera muitos outros recursos, como upload de conteúdo e aplicativos dinâmicos, tornando-o uma adição de desempenho a todos os aplicativos que servem tráfego pela Internet. O Lambda@Edge é um recurso do Amazon CloudFront que permite executar código mais perto dos usuários da sua carga de trabalho, o que melhora a performance e reduz a latência.

O Amazon Route 53 é um serviço web de DNS de nuvem altamente disponível e escalável. Ele é projetado visando oferecer aos desenvolvedores e empresas uma maneira extremamente confiável e econômica de rotear os usuários finais para aplicações de Internet convertendo nomes, como www.example.com, em endereços IP numéricos, como 192.168.2.1, que os computadores usam para se conectarem uns aos outros. O Route 53 é totalmente compatível com IPv6.

AWS Outposts foi projetado para workloads que precisam permanecer on-premises devido a requisitos de latência, quando a workload precisa ser executada sem problemas com o restante das outras workloads na AWS. Os AWS Outposts são racks de computação e armazenamento totalmente gerenciados e configuráveis criados com hardware projetado pela AWS, que permitem computação e armazenamento on-premises, e conectam-se sem problemas com a ampla matriz de serviços da AWS na nuvem.

As zonas locais da AWS são projetadas para executar workloads que exigem latência inferior a dez milissegundos, como renderização de vídeo e aplicações de desktop virtual com uso intenso de gráficos. As zonas locais permitem que você obtenha todos os benefícios de ter recursos de computação e armazenamento mais próximos dos usuários finais.

AWS Wavelength foi projetado para entregar aplicações de latência ultrabaixa para dispositivos 5G estendendo a infraestrutura, os serviços, as APIs e as ferramentas da AWS para as redes 5G. O Wavelength incorpora armazenamento e computação às redes 5G de provedores de telecomunicações para ajudar sua workload 5G, se ela exigir latência inferior a dez milissegundos, como dispositivos de IoT, streaming de jogos, veículos autônomos e produção de mídia ao vivo.

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

Antipadrões comuns:

- Você consolida todos os recursos da carga de trabalho em uma única localização geográfica.
- Você escolhe a região mais próxima ao seu local, mas não ao usuário final da workload.

Benefícios do estabelecimento desta prática recomendada: Verifique se sua rede está disponível sempre que quiser alcançar os clientes. O uso da rede global privada da AWS garante que seus clientes obtenham a experiência de menor latência ao implantar workloads nos locais mais próximos a eles.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

Reduzir a latência selecionando os locais corretos: identifique onde estão localizados os usuários e dados. Aproveite as Regiões da AWS, as zonas de disponibilidade, os grupos de posicionamento e os locais de borda para reduzir a latência.

Recursos

Documentos relacionados:

- Instâncias otimizadas para Amazon EBS
- Application Load Balancer
- Rede avançada do EC2 no Linux

- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Elastic Network Adapter (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- Transit Gateway
- Fazer a transição para o roteamento baseado em latência no Amazon Route 53
- VPC Endpoints
- Logs de fluxo da VPC

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1) (Conectividade com a AWS e arquiteturas de rede híbrida da AWS (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)

Exemplos relacionados:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluções de segurança escaláveis)
- AWS Networking Workshops

PERF05-BP07 Otimizar a configuração da rede com base em métricas

Use dados coletados e analisados para tomar decisões bem informadas sobre a otimização da configuração da rede. Meça o impacto dessas mudanças e use as medições de impacto para tomar decisões futuras.

Ative os logs de fluxo da VPC para todas as redes VPC usadas pela workload. O VPC Flow Logs é um recurso que permite capturar informações sobre o tráfego IP que entra e sai de interfaces de rede em sua VPC. O VPC Flow Logs ajuda você com várias tarefas, p. ex., com a solução de problemas relacionados à indisponibilidade de um tráfego específico para uma instância, o que, por sua vez, ajuda a diagnosticar regras excessivamente restritivas de grupos de segurança. Você pode usar os logs de fluxo como uma ferramenta de segurança para monitorar o tráfego que está chegando em instância, para criar o perfil do tráfego de rede e procurar comportamentos de tráfego anormais.

Use métricas de rede para fazer alterações na configuração de rede conforme a carga de trabalho evolui. É possível recriar as redes baseadas na nuvem rapidamente, portanto, é necessário evoluir sua arquitetura de rede ao longo do tempo para manter a eficiência da performance.

Antipadrões comuns:

- Você pressupõe que todos os problemas relacionados à performance são relacionados ao aplicativo.
- Você só testa a performance da rede a partir de um local próximo ao local em que implantou a carga de trabalho.

Benefícios do estabelecimento desta prática recomendada:para garantir que você esteja atendendo às métricas necessárias da workload, monitore as métricas de performance da rede. Você pode capturar informações sobre o tráfego IP de e para interfaces de rede em sua VPC e usar esses dados para adicionar novas otimizações ou implantar sua carga de trabalho em novas regiões geográficas.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

Ativar os logs de fluxo da VPC: os logs de fluxo da VPC permitem capturar informações sobre o tráfego IP de saída e de entrada das interfaces de rede em sua VPC. Os Logs de fluxo da VPC ajuda você com várias tarefas, como com a solução de problemas relacionados à indisponibilidade de um tráfego específico para uma instância, o que ajuda a diagnosticar regras excessivamente restritivas de grupos de segurança. Você pode usar os logs de fluxo como uma ferramenta de segurança para monitorar o tráfego que está chegando em instância, para criar o perfil do tráfego de rede e procurar comportamentos de tráfego anormais.

Ativar as métricas adequadas para as opções de rede: selecione as métricas de rede adequadas para a sua workload. Você pode habilitar métricas para o gateway NAT da VPC, gateways de trânsito e túneis VPN.

Recursos

Documentos relacionados:

- Instâncias otimizadas para Amazon EBS
- Application Load Balancer

- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Elastic Network Adapter (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- Transit Gateway
- Fazer a transição para o roteamento baseado em latência no Amazon Route 53
- VPC Endpoints
- Logs de fluxo da VPC
- Monitoramento de suas redes globais e principais com as métricas do Amazon CloudWatch
- Monitore continuamente o tráfego e os recursos da rede

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1) (Conectividade com a AWS e arquiteturas de rede híbrida da AWS (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)
- Monitoring and troubleshooting network traffic (Monitoramento e solução de problemas de tráfego de rede)
- Simplify Traffic Monitoring and Visibility with Amazon VPC Traffic Mirroring (Simplificar o monitoramento e a visibilidade do tráfego com a Amazon VPC Traffic Mirroring)

Exemplos relacionados:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluções de segurança escaláveis)
- AWS Networking Workshops
- Monitoramento de rede da AWS

Análise

Pergunta

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.

Práticas recomendadas

- PERF06-BP01 Manter-se atualizado sobre novos recursos e serviços
- PERF06-BP02 Definir um processo para melhorar a performance da workload
- PERF06-BP03 Evoluir a performance da workload ao longo do tempo

PERF06-BP01 Manter-se atualizado sobre novos recursos e serviços

Avalie maneiras de aumentar o desempenho conforme surgem novos serviços, padrões de design e ofertas de produtos. Determine quais deles poderiam aprimorar o desempenho ou aumentar a eficiência da workload por meio de avaliações, discussões internas ou análises externas.

Defina um processo para avaliar atualizações, novos recursos e serviços relevantes para sua workload. Por exemplo, criação de uma prova de conceito que use novas tecnologias ou consulta com um grupo interno. Ao testar novas ideias ou serviços, execute testes de desempenho para medir o impacto que eles têm sobre o desempenho da workload. Uso de infraestrutura como código (IaC) e uma cultura de DevOps para aproveitar a capacidade de testar novas ideias ou tecnologias frequentemente com mínimo custo ou risco.

Resultado desejado: você documentou o inventário dos componentes, o padrão de design e as características da workload. Você usa essa documentação para criar uma lista de assinaturas a fim de notificar sua equipe sobre atualizações de serviço, recursos e novos produtos. Você identificou as partes interessadas do componente que vão avaliar os lançamentos e fornecer uma recomendação para prioridade e impacto empresarial.

Antipadrões comuns:

- Você só analisa novas opções e serviços quando a workload não está atendendo aos requisitos de desempenho.
- Você pressupõe que as novas ofertas de produtos não sejam úteis para sua workload.
- Você sempre opta por criar em vez de comprar ao melhorar sua workload.

Benefícios de estabelecer esta prática recomendada: Ao considerar novos serviços ou ofertas de produtos, você pode melhorar o desempenho e a eficiência de sua workload, diminuir os custos de infraestrutura e reduzir o esforço exigido para manter seus serviços.

Nível de risco exposto se essa prática recomendada não for estabelecida: alto

Orientação para implementação

Defina um processo para avaliar atualizações, novos recursos e serviços da AWS. Por exemplo, a criação de provas de conceito que usam novas tecnologias. Ao testar novas ideias ou serviços, execute testes de performance para medir o impacto que eles têm sobre a eficiência ou a performance da workload. Aproveite a flexibilidade que você tem na AWS para testar novas ideias ou tecnologias frequentemente com custo ou risco mínimos.

Etapas da implementação

1. Documente as soluções da workload. Use a solução de banco de dados de gerenciamento de configurações (CMDB) para documentar seu inventário e categorizar serviços e dependências. Use ferramentas como o AWS Config para obter uma lista de todos os serviços na AWS que estão sendo usados por sua workload.
2. Use uma estratégia de marcação para documentar os proprietários de cada componente e categoria da workload. Por exemplo, se você estiver usando o Amazon RDS como solução de banco de dados, atribua e documente o administrador do banco de dados (DBA) como proprietário para avaliar e pesquisar novos serviços e atualizações.
3. Identifique novidades e atualize fontes relacionadas aos componentes da workload. No exemplo do Amazon RDS mencionado anteriormente, o proprietário da categoria deve assinar o blog What's New at AWS (Novidades da AWS) para os produtos que correspondem ao componente da workload. Você pode assinar o feed RSS ou gerenciar suas assinaturas de e-mail. Monitore atualizações do banco de dados do Amazon RDS que você utiliza, recursos introduzidos, instâncias lançadas e novos produtos, como o Amazon Aurora Serverless. Monitore blogs, produtos e fornecedores do setor do qual o componente depende.
4. Documente seu processo para avaliar atualizações e novos serviços. Forneça aos proprietários da categoria o tempo e o espaço necessários para pesquisar, testar, experimentar e validar atualizações e novos serviços. Consulte novamente os KPIs e requisitos empresariais documentados para ajudar a priorizar qual atualização trará um impacto positivo à empresa.

Nível de esforço do plano de implementação: Para estabelecer essa prática recomendada, é necessário estar ciente dos componentes atuais da workload, identificar os proprietários da categoria

e identificar as fontes das atualizações de serviço. Esse é um nível baixo de esforço para começar, mas é um processo contínuo que poderia evoluir e melhorar ao longo do tempo.

Recursos

Documentos relacionados:

- Blog da AWS
- Novidades da AWS

Vídeos relacionados:

- Canal AWS Events no YouTube
- Canal Online Tech Talks da AWS no YouTube
- Canal da Amazon Web Services no YouTube

Exemplos relacionados:

- AWS Github
- AWS Skill Builder

PERF06-BP02 Definir um processo para melhorar a performance da workload

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.

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

Antipadrões comuns:

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

Benefícios do estabelecimento desta prática recomendada: Ao definir seu processo para fazer alterações de arquitetura, você permite que os dados coletados influenciem o design da carga de trabalho ao longo do tempo.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

Identificar as restrições principais da performance da workload: documente as restrições da performance da workload para saber quais tipos de inovação podem aprimorar a performance da sua workload.

Recursos

Documentos relacionados:

- Blog da AWS
- Novidades da AWS

Vídeos relacionados:

- Canal AWS Events no YouTube
- Canal Online Tech Talks da AWS no YouTube
- Canal da Amazon Web Services no YouTube

Exemplos relacionados:

- AWS Github
- AWS Skill Builder

PERF06-BP03 Evoluir a performance da workload 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.

Use as informações coletadas ao avaliar novos serviços ou tecnologias para promover mudanças. Conforme seu negócio ou carga de trabalho muda, a performance também precisa mudar. Use dados coletados de suas métricas de carga de trabalho para avaliar áreas nas quais você pode obter

os maiores ganhos de eficiência ou performance e adote proativamente novos serviços e tecnologias para acompanhar a demanda.

Antipadrões comuns:

- Você pressupõe que sua arquitetura atual se tornará estática e nunca será atualizada ao longo do tempo.
- Você apresenta alterações de arquitetura ao longo do tempo sem justificativa de métrica.
- Você altera a arquitetura apenas porque todos os outros no setor a estão usando.

Benefícios do estabelecimento desta prática recomendada: Para aprimorar a performance e o custo da carga de trabalho, você deve avaliar todos os softwares e serviços disponíveis para determinar quais são os apropriados para sua carga de trabalho.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

Evoluir sua workload ao longo do tempo: use as informações coletadas ao avaliar novos serviços ou tecnologias para promover mudanças. A performance também precisa mudar conforme seu negócio ou carga de trabalho muda. Use dados coletados de suas métricas de carga de trabalho para avaliar áreas nas quais você pode obter os maiores ganhos de eficiência ou performance e adote proativamente novos serviços e tecnologias para acompanhar a demanda.

Recursos

Documentos relacionados:

- Blog da AWS
- Novidades da AWS

Vídeos relacionados:

- Canal AWS Events no YouTube
- Canal Online Tech Talks da AWS no YouTube
- Canal da Amazon Web Services no YouTube

Exemplos relacionados:

- AWS Github
- AWS Skill Builder

Monitoramento

Pergunta

- PERF 7 Como você monitora seus recursos para garantir que eles estejam apresentando boa performance?

PERF 7 Como você monitora seus recursos para garantir que eles estejam apresentando boa performance?

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.

Práticas recomendadas

- PERF07-BP01 Registrar métricas relacionadas à performance
- PERF07-BP02 Analisar as métricas quando ocorrem eventos ou incidentes
- PERF07-BP03 Estabelecer indicadores-chave de performance (KPIs) para medir a performance da workload
- PERF07-BP04 Usar o monitoramento para gerar notificações baseadas em alarme
- PERF07-BP05 Analisar as métricas em intervalos regulares
- PERF07-BP06 Monitorar e emitir alarmes proativamente

PERF07-BP01 Registrar métricas relacionadas à performance

Use um serviço de monitoramento e observação para registrar métricas relacionadas à performance. Os exemplos de métricas incluem registro de transações do banco de dados, consultas lentas, latência de E/S, throughput de solicitação HTTP, latência de serviço ou outros dados importantes.

Identifique as métricas de performance relevantes para sua carga de trabalho e registre-as. Esses dados são importantes para identificar quais componentes estão afetando a performance ou a eficiência geral da carga de trabalho.

Trabalhando no sentido contrário à experiência do cliente, identifique as métricas relevantes. Para cada métrica, identifique o alvo, a abordagem de medição e a prioridade. Use esses dados para criar alarmes e notificações visando abordar proativamente problemas relacionados à performance.

Antipadrões comuns:

- Você só monitora métricas no nível do sistema operacional para obter informações sobre sua carga de trabalho.
- Você arquiteta suas necessidades de computação para os requisitos de pico de carga de trabalho.

Benefícios do estabelecimento desta prática recomendada: Para otimizar a performance e a utilização de recursos, você precisa de uma visão operacional unificada dos seus indicadores-chave de performance. Você pode criar painéis e executar matemática de métricas em seus dados para obter insights operacionais e de utilização.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

Identifique as métricas de performance relevantes para sua carga de trabalho e registre-as. Esses dados ajudam a identificar quais componentes estão afetando a performance geral ou a eficiência da carga de trabalho.

Identificar métricas de performance: use a experiência do cliente para identificar as métricas mais importantes. Para cada métrica, identifique o alvo, a abordagem de medição e a prioridade. Use esses pontos de dados para criar alarmes e notificações visando abordar proativamente problemas relacionados à performance.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Coletar métricas e logs das instâncias do Amazon EC2 e de servidores on-premises com o agente do CloudWatch
- Publicar métricas personalizadas
- Monitoramento, registro em log e performance dos parceiros da APN
- Documentação do X-Ray

- Amazon CloudWatch RUM

Vídeos relacionados:

- Cut through the chaos: Gain operational visibility and insight (MGT301-R1)
- Application Performance Management na AWS
- Build a monitoring plan

Exemplos relacionados:

- Nível 100: monitoramento com os painéis do CloudWatch
- Nível 100: monitoramento das instâncias do Windows do EC2 com os painéis do CloudWatch
- Nível 100: monitoramento de uma instância do Amazon Linux EC2 com os painéis do CloudWatch

PERF07-BP02 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.

Ao escrever histórias de usuário importantes para sua arquitetura, inclua requisitos de performance, como especificar a rapidez com que cada história de usuário importante deve ser executada. Para essas histórias críticas, implemente jornadas de usuários em roteiros adicionais, para saber como elas se comportam em relação aos seus requisitos.

Antipadrões comuns:

- Você pressupõe que os eventos de performance são problemas pontuais e que estão relacionados apenas a anomalias.
- Você só avalia métricas de performance existentes ao responder a eventos de performance.

Benefícios do estabelecimento desta prática recomendada: Ao determinar se sua workload está operando nos níveis esperados, responda aos eventos de performance coletando dados de métrica adicionais para análise. Esses dados são usados para compreender o impacto do evento de performance e sugerir alterações para melhorar a performance da carga de trabalho.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

Priorizar as preocupações sobre experiência de histórias críticas de usuários: ao escrever histórias críticas de usuários para sua arquitetura, inclua requisitos de performance, como especificar a rapidez com que cada história crítica deve ser executada. Para essas histórias essenciais, implemente jornadas de usuário em roteiros adicionais, de modo que você conheça a performance delas em relação aos seus requisitos.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Amazon CloudWatch Synthetics
- Monitoramento, registro em log e performance dos parceiros da APN
- Documentação do X-Ray

Vídeos relacionados:

- Cut through the chaos: Gain operational visibility and insight (MGT301-R1)
- Optimize applications through Amazon CloudWatch RUM (Otimizar as aplicações por meio do Amazon CloudWatch RUM)
- Demo of Amazon CloudWatch Synthetics (Demonstração do Amazon CloudWatch Synthetics)

Exemplos relacionados:

- Medição do tempo de carga da página com o Amazon CloudWatch Synthetics
- Cliente da web do Amazon CloudWatch RUM

PERF07-BP03 Estabelecer indicadores-chave de performance (KPIs) para medir a performance da workload

Identifique os KPIs que medem a performance da workload de forma quantitativa e qualitativa. Os KPIs ajudam a medir a integridade de uma workload quando relacionada a uma meta dos negócios. Os KPIs permitem que as empresas e as equipes de engenharia alinhem a medição das metas e estratégias de como isso é combinado para produzir resultados para os negócios. Os KPIs devem

ser revisitados quando mudam as metas, as estratégias da empresa ou os requisitos dos usuários finais.

Por exemplo, a workload de um site pode usar o tempo de carregamento da página como uma indicação de performance geral. Essa métrica seria um dos vários pontos de dados que medem a experiência de um usuário final. Além de identificar os limites do tempo de carregamento da página, documente o resultado esperado ou o risco da empresa se a performance não for atendida. Um tempo longo de carregamento da página afeta diretamente os usuários finais, reduz a taxa da experiência dos usuários e pode resultar em perda de clientes. Ao definir os limites de seus KPIs, combine os testes comparativos do setor e as expectativas dos seus usuários finais. Por exemplo, se o teste comparativo do setor atual for o carregamento de uma página da web em dois segundos, mas seus usuários finais esperarem que uma página da web seja carregada em um segundo, você deverá considerar os dois pontos de dados ao estabelecer o KPI. Outro exemplo de um KPI pode ser focalizado no atendimento das necessidades internas de performance. Um limite de KPI pode ser estabelecido para a geração de relatórios de vendas em um dia útil depois da geração dos dados da produção. Esses relatórios podem afetar diretamente as decisões diárias e os resultados da empresa.

Resultado desejado: O estabelecimento de KPIs envolve diferentes departamentos e partes interessadas. Sua equipe deve avaliar as KPIs da sua workload usando dados detalhados em tempo real e dados históricos para referência, e criar painéis que calculem as métricas em seus dados de KPI para derivar insights operacionais e de utilização. Os KPIs devem ser documentados para explicar os KPIs concordados e os limites compatíveis com as metas e estratégias da empresa bem como mapeados para as métricas monitoradas. Os KPIS são requisitos da performance identificados, que devem ser revistos intencionalmente e compartilhados e compreendidos frequentemente com todas as equipes. Os riscos e as compensações devem ser claramente identificados e compreendidos, para o caso dos limites dos KPIs não serem atendidos.

Antipadrões comuns:

- Você só monitora as métricas em nível do sistema para obter insight de sua workload e não compreende aos impactos dessas métricas nos negócios.
- Você pressupõe que seus KPIs já estejam publicados como dados de métricas padrão.
- Você define os KPIs, mas não os compartilha com todas as equipes.
- Você não define um KPI quantitativo e mensurável.
- Você não alinha os KPIs com as metas e as estratégias dos negócios.

Benefícios do estabelecimento desta prática recomendada: Identificação das métricas específicas que representam a integridade da workload para alinhar as equipes com suas prioridades e a definição bem-sucedida dos resultados da empresa. O compartilhamento dessas métricas com todos os departamento fornece visibilidade e alinhamento dos limites, das expectativas e do impacto nos negócios.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

Todos os departamentos e equipes da empresa afetados pela integridade da workload devem contribuir para a definição dos KPIs. Uma única pessoa deve orientar a colaboração, os cronogramas, a documentação e as informações relacionadas às KPIs de uma organização. Esse proprietário único compartilhará as metas e as estratégias da empresa com frequência, e atribuirá tarefas às partes interessadas da empresa para criarem KPIs em seus respectivos departamentos. Depois da definição dos KPIs, a equipe de operações ajudará a definir as métricas que apoiarão e informarão o sucesso dos diferentes KPIs. Os KPIs só serão eficazes se todos os membros da equipe que oferece suporte a uma workload tiverem ciência os KPIs.

Etapas da implementação

1. Identificar e documentar as partes interessadas da empresa.
2. Identificar as metas e as estratégias da empresa.
3. Analisar os KPIs comuns do setor que se alinham com as metas e estratégias da empresa.
4. Analisar as expectativas dos usuários finais em relação à sua workload.
5. Definir e documentar os KPIs que oferecem suporte às metas e às estratégias da empresa.
6. Identificar e documentar as estratégias de compensação para atender aos KPIs.
7. Identificar e documentar as métricas que fornecerão informações dos KPIs.
8. Identificar e documentar os limites dos KPIs por nível de gravidade ou de alarme.
9. Identificar e documentar o risco e o impacto no caso e um KPI não ser atendido.
10. Identificar a frequência de revisão por KPI.
11. Comunicar a documentação dos KPIs a todas as equipes que oferecem suporte à workload.

Nível de esforço das orientações de implementação: A definição e a comunicação dos KPIs é uma baixo quantidade de trabalho. Normalmente, isso pode ser feito em algumas semanas, em reuniões com as partes interessadas, analisando as metas, as estratégias e as métricas da workload.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Monitoramento, registro em log e performance dos parceiros da APN
- Documentação do X-Ray
- Uso de painéis do Amazon CloudWatch
- KPIs do Amazon QuickSight

Vídeos relacionados:

- AWS re:Invent 2019: Scaling up to your first 10 million users (ARC211-R) (Aumentar a escala verticalmente para atingir seus primeiros dez milhões de usuários (ARC211-R))
- Cut through the chaos: Gain operational visibility and insight (MGT301-R1)
- Build a monitoring plan

Exemplos relacionados:

- Criação de um painel com o Amazon QuickSight

PERF07-BP04 Usar o 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.

O Amazon CloudWatch pode coletar métricas nos recursos da sua arquitetura. Você também pode coletar e publicar métricas personalizadas para descobrir métricas de negócio ou derivadas. Use o CloudWatch ou um serviço de monitoramento de terceiros para definir alarmes que indiquem quando há uma violação de limites. Os alarmes sinalizam que uma métrica está fora dos limites esperados.

Antipadrões comuns:

- Você depende das equipes para observar métricas e reagir quando elas percebem um problema.
- Você depende apenas de runbooks operacionais, quando fluxos de trabalho de tecnologia sem servidor poderiam ser acionados para realizar a mesma tarefa.

Benefícios do estabelecimento desta prática recomendada: Você pode definir alarmes e automatizar ações com base em limites predefinidos ou em algoritmos de Machine Learning que identificam comportamento anormal em suas métricas. Esses mesmos alarmes também podem acionar fluxos de trabalho de tecnologia sem servidor, que podem modificar características de performance da sua workload (por exemplo, aumento da capacidade computacional, alteração da configuração do banco de dados).

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

Monitorar métricas: o Amazon CloudWatch pode coletar métricas entre os recursos da sua arquitetura. Você pode coletar e publicar métricas personalizadas para descobrir métricas de negócio ou derivadas. Use o CloudWatch ou um serviço de monitoramento de terceiros para definir alarmes que indiquem quando os limites forem excedidos.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Monitoramento, registro em log e performance dos parceiros da APN
- Documentação do X-Ray
- Using Alarms and Alarm Actions in CloudWatch (Usar alarmes e ações de alarmes no CloudWatch)

Vídeos relacionados:

- AWS re:Invent 2019: Scaling up to your first 10 million users (ARC211-R) (Aumentar a escala verticalmente para atingir seus primeiros dez milhões de usuários (ARC211-R))
- Cut through the chaos: Gain operational visibility and insight (MGT301-R1)
- Build a monitoring plan
- Using AWS Lambda with Amazon CloudWatch Events (Usar o Amazon Lambda com o Amazon CloudWatch)

Exemplos relacionados:

- Cloudwatch Logs Customize Alarms (Os logs do CloudWatch personalizam os alarmes)

PERF07-BP05 Analisar as métricas em intervalos regulares

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

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

Antipadrões comuns:

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

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

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

Melhorar constantemente a coleta e o monitoramento das métricas: como parte das respostas a incidentes ou eventos, avalie as métricas que foram úteis ao resolver o problema e quais métricas, que não estão sendo acompanhadas, ajudariam a resolver o problema; Use este método para aprimorar a qualidade das métricas coletadas, de modo que você possa prevenir ou resolver incidentes futuros mais rapidamente.

Recursos

Documentos relacionados:

- Documentação do CloudWatch

- Coletar métricas e logs das instâncias do Amazon EC2 e de servidores on-premises com o agente do CloudWatch
- Monitoramento, registro em log e performance dos parceiros da APN
- Documentação do X-Ray

Vídeos relacionados:

- Cut through the chaos: Gain operational visibility and insight (MGT301-R1)
- Application Performance Management on AWS (Gerenciamento da performance de aplicações na AWS)
- Build a monitoring plan

Exemplos relacionados:

- Criação de um painel com o Amazon QuickSight
- Level 100: Monitoring with CloudWatch Dashboards (Nível 100: monitoramento com os painéis do CloudWatch)

PERF07-BP06 Monitorar e emitir alarmes proativamente

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.

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

Antipadrões comuns:

- Você só permite que a equipe de operações faça alterações operacionais na carga de trabalho.

- Você permite todos os filtros de alarmes para a equipe de operações, sem correção proativa.

Benefícios do estabelecimento desta prática recomendada: A correção proativa de ações de alarme permite que a equipe de suporte se concentre nos itens que não são acionáveis automaticamente. Isso garante que a equipe de operações não seja sobrecarregada por todos os alarmes e, em vez disso, se concentre apenas em alarmes críticos.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

Monitorar a performance durante as operações: implemente processos que forneçam visibilidade da performance à medida que sua workload evolui. Crie painéis de monitoramento e estabeleça uma linha de base para as expectativas de performance.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Monitoramento, registro em log e performance dos parceiros da APN
- Documentação do X-Ray
- Using Alarms and Alarm Actions in CloudWatch (Usar alarmes e ações de alarmes no CloudWatch)

Vídeos relacionados:

- Cut through the chaos: Gain operational visibility and insight (MGT301-R1)
- Application Performance Management on AWS (Gerenciamento da performance de aplicações na AWS)
- Build a monitoring plan
- Using AWS Lambda with Amazon CloudWatch Events (Usar o Amazon Lambda com o Amazon CloudWatch)

Exemplos relacionados:

- Cloudwatch Logs Customize Alarms (Os logs do CloudWatch personalizam os alarmes)

Concessões

Pergunta

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

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.

Práticas recomendadas

- PERF08-BP01 Compreender as áreas em que o desempenho é mais importante
- PERF08-BP02 Saber mais sobre serviços e padrões de design
- PERF08-BP03 Identificar como as compensações afetam os clientes e a eficiência
- PERF08-BP04 Medir o impacto das melhorias na performance
- PERF08-BP05 Usar várias estratégias relacionadas à performance

PERF08-BP01 Compreender as áreas em que o desempenho é mais importante

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

Resultado desejado: aumentar a eficiência do desempenho entendendo sua arquitetura, os padrões de tráfego e os padrões de acesso aos dados, além de identificar os tempos de latência e processamento. Identificar possíveis gargalos que possam afetar a experiência do cliente com o crescimento da workload. Ao identificar essas áreas, veja qual solução você pode implantar para remover essas preocupações com o desempenho.

Antipadrões comuns:

- Você presume que as métricas de computação comuns, como `CPUUtilization` ou pressão de memória são suficientes para capturar os problemas de desempenho.
- Você só usa as métricas comuns registradas pelo software de monitoramento selecionado.

- Você só revisa as métricas quando há um problema.

Benefícios de estabelecer esta prática recomendada: Compreender áreas críticas de desempenho ajuda os proprietários de workloads a monitorar KPIs e priorizar melhorias de alto impacto.

Nível de risco exposto se essa prática recomendada não for estabelecida: alto

Orientação para implementação

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

Etapas da implementação

1. Configure um monitoramento completo para capturar todos os componentes e as métricas da workload.
   - Use o Monitoramento de usuários reais (RUM) do Amazon CloudWatch para capturar as métricas de desempenho da aplicação de sessões de front-end e do lado do cliente de usuários reais.
   - Configure o AWS X-Ray para rastrear o tráfego por meio das camadas de aplicação e identificar a latência entre componentes e dependências. Use os mapas do serviço X-Ray para ver os relacionamentos e a latência entre os componentes da workload.
   - Use o Insights de Performance do Amazon Relational Database Service para ver as métricas de desempenho do banco de dados e identificar melhorias de desempenho.
   - Use o Monitoramento avançado do Amazon RDS para ver métricas de desempenho do SO do banco de dados.
   - Colete métricas do CloudWatch por componente e serviço da workload e identifique quais métricas afetam a eficiência do desempenho.
   - Configure o Amazon DevOps Guru para obter recomendações e insights de desempenho adicionais.
2. Realize testes para gerar métricas, identificar padrões de tráfego, gargalos e áreas de desempenho críticas.
   - Configure o Canários sintéticos do CloudWatch para imitar as atividades do usuário no navegador de forma programática usando trabalhos `cron` ou expressões de avaliação para gerar métricas consistentes ao longo do tempo.

    - Use a solução de Testes de carga distribuída da AWS para gerar tráfego de pico ou testar a workload na taxa de crescimento esperada.
3. Avalie as métricas e a telemetria para identificar as áreas de desempenho críticas. Avalie essas áreas com sua equipe para discutir sobre o monitoramento e as soluções visando evitar gargalos.
4. Experimente melhorias de desempenho e meça essas alterações com dados.
    - Use o CloudWatch Evidently para testar novas melhorias e o impacto do desempenho na workload.

Nível de esforço do plano de implementação: Para estabelecer essa prática recomendada, é necessário analisar suas métricas completas e estar ciente do desempenho atual da workload. Esse é um nível moderado de esforço para configurar o monitoramento completo e identificar as áreas de desempenho críticas.

Recursos

Documentos relacionados:

- Amazon Builders' Library
- Documentação do X-Ray
- Amazon CloudWatch RUM
- Amazon DevOps Guru
- CloudWatch RUM e X-Ray

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- Demo of Amazon CloudWatch Synthetics (Demonstração do Amazon CloudWatch Synthetics)

Exemplos relacionados:

- Measure page load time with Amazon CloudWatch Synthetics (Medição do tempo de carga da página com o Amazon CloudWatch Synthetics)
- Amazon CloudWatch RUM Web Client (Cliente da web do Amazon CloudWatch RUM)
- X-Ray SDK para Node.js
- X-Ray SDK para Python

- X-Ray SDK para Java
- X-Ray SDK para .Net
- X-Ray SDK para Ruby
- Daemon do X-Ray
- Testes de carga distribuída na AWS

PERF08-BP02 Saber mais 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 colocada nos sistemas de banco de dados. No entanto, o armazenamento em cache pode introduzir uma eventual consistência e requer esforço de engenharia para ser implementado de acordo com os requisitos de negócios e as expectativas dos clientes.

Resultado desejado: A pesquisa de padrões de design levará você a escolher um design de arquitetura que oferecerá suporte ao sistema com melhor performance. Saiba quais opções de configuração de performance estão disponíveis e como elas poderiam afetar a carga de trabalho. A otimização da performance de sua workload exige entender como essas opções interagem com sua arquitetura e o impacto que elas terão sobre a performance medida e a performance percebida pelos usuários.

Antipadrões comuns:

- Você pressupõe que todas as estratégias tradicionais de performance de cargas de trabalho de TI são mais adequadas para cargas de trabalho na nuvem.
- Você cria e gerencia soluções de armazenamento em cache em vez de usar serviços gerenciados.
- Você usa o mesmo padrão de design para todas as workloads sem avaliar qual padrão melhoraria a performance da workload.

Benefícios do estabelecimento desta prática recomendada: Ao selecionar o padrão de design e os serviços certos para sua workload, você vai otimizar a performance, melhorar a excelência operacional e aumentar a confiabilidade. O padrão de design ideal vai atender às características de sua workload atual e ajudar você a dimensionar para crescimento ou alterações futuras.

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

Orientação de implementação

Saiba quais opções de configuração de performance estão disponíveis e como elas poderiam afetar a carga de trabalho. A otimização da performance de sua carga de trabalho exige entender como essas opções interagem com sua arquitetura e o impacto que elas terão sobre a performance medida e a performance percebida pelo usuário.

Etapas da implementação:

1. Avalie e revise os padrões de design que melhorariam a performance de sua workload.
    a. O Amazon Builders' Library fornece uma descrição detalhada de como a Amazon cria e opera tecnologias. Estes artigos são escritos por engenheiros seniores da Amazon e abordam temas sobre arquitetura, entrega de software e operações.
    b. Biblioteca de Soluções da AWS é um conjunto de soluções prontas para implantar que reúnem serviços, código e configurações. Essas soluções foram criadas pela AWS e por parceiros da AWS com base em casos de uso comuns e padrões de design agrupados por setor e tipo de workload. Por exemplo, você pode configurar uma solução de testes de carga distribuída para a workload.
    c. Centro de Arquitetura da AWS fornece diagramas de arquitetura de referência agrupados por padrão de design, tipo de conteúdo e tecnologia.
    d. Amostras da AWS é um repositório do GitHub repleto de exemplos práticos para ajudar você a explorar padrões de arquitetura comuns, soluções e serviços. É atualizado frequentemente com os serviços e exemplos mais recentes.
2. Melhore sua workload para modelar os padrões de design selecionados e use os serviços e as opções de configuração de serviços para melhorar a performance de sua workload.
    a. Treine sua equipe interna com os recursos disponíveis no AWS Skills Guild.
    b. Use a ferramenta de recomendações do AWS Partner Network para oferecer experiência com rapidez e para escalar sua capacidade de implementar melhorias.

Nível de esforço do plano de implementação: Para estabelecer esta prática recomendada, você deve conhecer os padrões de design e os serviços capazes de ajudar a melhorar a performance de sua workload. Depois de avaliar os padrões de design, a implementação desses padrões representa um esforço de nível alto .

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS
- Amazon Builders' Library
- Como usar o descarte de carga para evitar sobrecarga
- Desafios e estratégias de armazenamento em cache

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- This is My Architecture (Esta é a minha arquitetura)

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS

PERF08-BP03 Identificar como as compensações 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.

Identifique áreas de baixa performance em seu sistema por meio de métricas e monitoramento. Determine como você pode promover aprimoramentos, quais concessões esses aprimoramentos exigem e como elas afetam o sistema e a experiência do usuário. Por exemplo, a implementação de armazenamento de dados em cache pode ajudar a aprimorar drasticamente a performance, mas requer uma estratégia clara de como e quando atualizar ou invalidar dados em cache a fim de prevenir comportamentos incorretos do sistema.

Antipadrões comuns:

- Você pressupõe que todos os ganhos de performance devem ser implementados, mesmo que haja compensações para implementação, como consistência eventual.

- Você só avalia alterações nas cargas de trabalho quando um problema de performance atinge um ponto crítico.

Benefícios do estabelecimento desta prática recomendada: Ao avaliar possíveis melhorias relacionadas à performance, você deve decidir se as compensações para as alterações são consistentes com os requisitos da carga de trabalho. Em alguns casos, pode ser necessário implementar controles adicionais para compensar as compensações.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Alto

Orientações para a implementação

Identificar compensações: use métricas e monitoramento para identificar áreas com performance insatisfatória em seu sistema. Determine como fazer melhorias e como as compensações afetarão o sistema e a experiência do usuário. Por exemplo, a implementação de armazenamento de dados em cache pode ajudar a aprimorar drasticamente a performance, mas requer uma estratégia clara de como e quando atualizar ou invalidar dados em cache a fim de evitar comportamentos incorretos do sistema.

Recursos

Documentos relacionados:

- Amazon Builders' Library
- KPIs do Amazon QuickSight
- Amazon CloudWatch RUM
- Documentação do X-Ray

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- Build a monitoring plan
- Optimize applications through Amazon CloudWatch RUM (Otimizar as aplicações por meio do Amazon CloudWatch RUM)
- Demo of Amazon CloudWatch Synthetics (Demonstração do Amazon CloudWatch Synthetics)

Exemplos relacionados:

- Medição do tempo de carga da página com o Amazon CloudWatch Synthetics
- Cliente da web do Amazon CloudWatch RUM

PERF08-BP04 Medir o impacto das melhorias na 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.

Um sistema Well-Architected usa uma combinação de estratégias relacionadas à performance. Determine que estratégia terá o maior impacto positivo sobre um dado hotspot ou gargalo. Por exemplo, a fragmentação de dados em vários sistemas de bancos de dados relacionais poderia aumentar o throughput geral e ao mesmo tempo manter o suporte para transações e, dentro de cada fragmento, o armazenamento em cache pode ajudar a reduzir a carga.

Antipadrões comuns:

- Você implanta e gerencia manualmente tecnologias que estão disponíveis como serviços gerenciados.
- Você se concentra em apenas um componente, como redes, quando vários componentes podem ser usados para aumentar a performance da carga de trabalho.
- Você conta com o feedback e as percepções de clientes como seu único teste comparativo.

Benefícios do estabelecimento desta prática recomendada: Para implementar estratégias de performance, selecione vários serviços e recursos que, juntos, permitirão atender aos requisitos de performance da workload.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

Um sistema bem arquitetado usa uma combinação de estratégias relacionadas à performance. Determine qual estratégia terá o maior impacto positivo sobre um dado hotspot ou gargalo. Por exemplo, a fragmentação de dados em vários sistemas de bancos de dados relacionais poderia aumentar o throughput geral e ao mesmo tempo manter o suporte para transações e, dentro de cada fragmento, o armazenamento em cache pode ajudar a reduzir a carga.

Recursos

Documentos relacionados:

- Amazon Builders' Library
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Testes de carga distribuída na AWS

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- Optimize applications through Amazon CloudWatch RUM (Otimizar as aplicações por meio do Amazon CloudWatch RUM)
- Demo of Amazon CloudWatch Synthetics (Demonstração do Amazon CloudWatch Synthetics)

Exemplos relacionados:

- Medição do tempo de carga da página com o Amazon CloudWatch Synthetics
- Cliente da web do Amazon CloudWatch RUM
- Testes de carga distribuída na AWS

PERF08-BP05 Usar 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 leitura, 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.

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.

Antipadrões comuns:

- Você pressupõe que a performance da carga de trabalho seja adequada se os clientes não estiverem reclamando.
- Você só coleta dados sobre a performance depois de fazer alterações relacionadas a ela.

Benefícios do estabelecimento desta prática recomendada: Você precisa de uma visão operacional unificada, dados granulares em tempo real e uma referência histórica para otimizar a performance e a utilização de recursos. Você pode criar painéis e executar matemática de métricas em seus dados para obter insights operacionais e de utilização para suas cargas de trabalho à medida que elas mudam ao longo do tempo.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

Usar uma abordagem orientada por dados para evoluir a arquitetura: conforme você altera a workload, 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

Documentos relacionados:

- Amazon Builders' Library
- Melhores práticas para a implementação do Amazon ElastiCache
- Armazenamento em cache de banco de dados da AWS
- Amazon CloudWatch RUM
- Testes de carga distribuída na AWS

Vídeos relacionados:

- Introducing The Amazon Builders' Library (DOP328)
- AWS purpose-built databases (DAT209-L) (Bancos de dados com propósito específico da AWS (DAT209-L))
- Optimize applications through Amazon CloudWatch RUM

Exemplos relacionados:

- Medição do tempo de carga da página com o Amazon CloudWatch Synthetics
- Cliente da web do Amazon CloudWatch RUM
- Testes de carga distribuída na AWS

Otimização de custos

Tópicos

Pratique o gerenciamento financeiro na nuvem

Pergunta

COST 1 Como implementar o gerenciamento financeiro na nuvem?

A implementação do gerenciamento financeiro na nuvem possibilita que as organizações obtenham valor empresarial e sucesso financeiro à medida que elas otimizam os custos e o uso e escalam na AWS.

Práticas recomendadas

- COST01-BP01 Estabelecer uma função de otimização de custos
- COST01-BP02 Estabelecer uma parceria entre finanças e tecnologia
- COST01-BP03 Estabeleça orçamentos e previsões de nuvem
- COST01-BP04 Implemente o reconhecimento de custos em seus processos organizacionais
- COST01-BP05 Relatar e notificar sobre a otimização de custos
- COST01-BP06 Monitore custos proativamente

- COST01-BP07 Manter-se atualizado com os novos lançamentos de serviços

COST01-BP01 Estabelecer uma função de otimização de custos

Crie uma equipe (Escritório de Negócios na Nuvem ou Centro de Excelência da Nuvem) 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.

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

Orientação de implementação

Instaure uma equipe de Escritório de Negócios na Nuvem (CBO) ou Centro de Excelência da Nuvem (CCoE) responsável por estabelecer e manter uma cultura de reconhecimento de custos de computação em nuvem. Em toda a organização, tal função pode ser exercida por qualquer pessoa ou equipe existente, ou por uma nova equipe com as principais partes interessadas em finanças, tecnologia e organização.

A função (individual ou equipe) prioriza e gasta a porcentagem necessária de seu tempo em atividades de gerenciamento e otimização de custos. Para uma organização pequena, a função pode gastar uma porcentagem de tempo menor em comparação com uma função de tempo integral para uma empresa maior.

A função exige uma abordagem multidisciplinar, com recursos de gerenciamento de projetos, ciência de dados, análise financeira e desenvolvimento de software/infraestrutura. A função pode aumentar as eficiências de workloads ao executar otimizações de custo em três domínios diferentes:

- Centralizado: por meio de equipes designadas, como operações financeiras, otimização de custo, CBO ou CCoE, os clientes podem projetar e implementar mecanismos de governança e impulsionar práticas recomendadas em toda a empresa.
- Descentralizado: influenciando equipes de tecnologia a executar otimizações.
- Híbrido: uma combinação de equipes centralizadas e descentralizadas trabalhando em conjunto para executar otimizações de custo.

A função pode ser medida em relação à sua capacidade de executar e entregar em relação às metas de otimização de custos (por exemplo, métricas de eficiência da carga de trabalho).

Você deve garantir patrocínio executivo para que essa função estabeleça mudanças, o que é um fator de sucesso fundamental. O patrocinador é considerado defensor do consumo de nuvem

econômico e oferece suporte ao escalonamento para a função para garantir que as atividades de otimização de custos sejam tratadas com o nível de prioridade definido pela organização. Caso contrário, a orientação será ignorada e as oportunidades de economia de custo não serão priorizadas. Juntos, o patrocinador e a função garantem que sua organização consuma a nuvem com eficiência e continue a oferecer valor empresarial.

Se você tem um plano Business, Enterprise-On-Ramp ou Enterprise Support e precisa de ajuda para elaborar essa equipe ou função, entre em contato com especialistas do gerenciamento financeiro na nuvem (CFM) por meio de sua equipe de conta.

Etapas da implementação

- Defina os membros principais: Você precisa garantir que todas as partes relevantes da sua organização contribuam e tenham uma participação no gerenciamento de custos. As equipes comuns dentro das organizações geralmente incluem: finanças, proprietários de aplicações ou produtos, gerenciamento e equipes técnicas (DevOps). Algumas estão envolvidas em tempo integral (finanças, técnicas), outras periodicamente, conforme necessário. Indivíduos ou equipes que executam CFM geralmente precisam dos seguintes conjuntos de habilidades:
  - Habilidades de desenvolvimento de software: quando está ocorrendo desenvolvimento de scripts e automação.
  - Habilidades de engenharia de infraestrutura: para implantação de scripts ou automação e para entender como os serviços e recursos são provisionados.
  - Perspicácia operacional: CFM se trata de operar na nuvem de maneira eficiente por meio de medição, monitoramento, modificação, planejamento e dimensionamento do uso eficiente da nuvem.
- Definir metas e métricas: a função precisa agregar valor à organização de diferentes formas. Esses objetivos são definidos e evoluem continuamente com a organização. As atividades comuns incluem: criação e execução de programas educacionais sobre otimização de custos em toda a organização, desenvolvimento de padrões em toda a organização, como monitoramento e geração de relatórios para otimização de custos e definição de metas de workload sobre otimização. Essa função também precisa informar regularmente a organização sobre o recurso de otimização de custos das organizações.

  Você pode definir indicadores-chave de desempenho (KPIs) baseados em valor. Os KPIs podem ser baseados em custo ou baseados em valor. Ao definir os KPIs, você pode calcular o custo esperado em termos de eficiência e o resultado comercial esperado. KPIs baseados em valor vinculam métricas de uso e custo a direcionadores de valor comercial e nos ajudam a racionalizar

mudanças em nossos gastos na AWS. O primeiro passo para derivar KPIs baseados em valor é trabalhar em conjunto, em toda a organização, para selecionar e concordar sobre um conjunto padrão de KPIs.

- Estabelecer um ritmo regular: o grupo (equipes financeira, comercial e de tecnologia) devem se reunir regularmente para analisar metas e métricas. Um ritmo típico envolve analisar o estado da organização, todos os programas em execução no momento e as métricas financeiras e de otimização gerais. Em seguida, as principais workloads são relatadas em mais detalhes.

  Durante essas reuniões regulares, você pode analisar a eficiência (custo) da workload e o resultado comercial. Por exemplo, um aumento de 20% no custo de uma workload pode ser consequência de um aumento do uso pelos clientes. Neste caso, esse aumento de 20% no custo pode ser interpretado como um investimento. Essas chamadas regulares podem ajudar as equipes a identificarem KPIs baseados em valor que ofereçam propósito para toda a organização.

Recursos

Documentos relacionados:

- Blog de CCoE da AWS
- Criar um Escritório de Negócios na Nuvem
- CCoE: Centro de Excelência da Nuvem

Vídeos relacionados:

- Vanguard CCOE Success Story (História de sucesso de CCoE de vanguarda)

Exemplos relacionados:

- Usar um Centro de Excelência da Nuvem (CCoE) para transformar toda a empresa
- Criar um CCoE para transformar toda a empresa
- 7 obstáculos que devem ser evitados ao criar um CCoE

COST01-BP02 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.

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

Orientação de implementação

as equipes de tecnologia inovam mais rapidamente na nuvem devido à redução dos ciclos de implantação de aprovação, aquisição e infraestrutura. Isso pode ser um ajuste para organizações financeiras anteriormente usadas para executar processos demorados e com uso intensivo de recursos para aquisição e implantação de capital em ambientes de datacenter no local, além de alocação de custos apenas na aprovação do projeto.

Do ponto de vista da organização financeira e de aquisição, o processo de definição orçamentária, solicitações de capital, aprovações, aquisição e instalação de infraestrutura física é algo que levou décadas para ser aprendido e padronizado:

- Equipes de engenharia ou TI costumam ser os solicitantes
- Várias equipes financeiras atuam como aprovadores e compradores
- Equipes de operação estendem, acumulam e disponibilizam infraestrutura pronta para ser usada

Com a adoção da nuvem, a aquisição e o consumo de infraestrutura deixaram de estar vinculados a uma série de dependências. No modelo de nuvem, as equipes de tecnologia e produto deixam de ser simples desenvolvedoras, passando a ser operadoras e proprietárias de seus produtos, responsáveis pela maioria das atividades historicamente associadas às equipes financeiras e de operações, incluindo aquisição e implantação.

Basta uma conta de usuário e o conjunto adequado de permissões para provisionar recursos na nuvem. Também é isso que reduz o risco financeiro e de TI, o que significa que as equipes estão sempre a poucos cliques ou chamadas de API de encerrar recursos ociosos ou desnecessários na nuvem. Também é isso que permite que as equipes de tecnologia inovem com mais rapidez: a agilidade e capacidade de aplicar e derrubar experimentos. Embora a natureza variável do consumo

na nuvem possa afetar a previsibilidade do ponto de vista de previsão e definição orçamentária, a nuvem oferece às organizações a capacidade de reduzir o custo de provisionamento em excesso, além de reduzir o custo de oportunidade associado ao subprovisionamento conservador.

Estabelecer uma parceria entre as principais partes interessadas em finanças e tecnologia para criar uma compreensão compartilhada dos objetivos organizacionais e desenvolver mecanismos para obter sucesso financeiro no modelo de gastos variáveis da computação em nuvem. As equipes relevantes da sua organização devem estar envolvidas em discussões de custo e uso em todas as fases da jornada para a nuvem, incluindo:

- Líderes financeiros: CFOs, controladores financeiros, planejadores financeiros, analistas de negócios, aquisições, sourcing e contas a pagar devem compreender o modelo de nuvem de consumo, as opções de compra e o processo de faturamento mensal. O departamento financeiro precisa se unir às equipes de tecnologia para criar e socializar uma narrativa de valor de TI, ajudando as equipes comerciais a entender como o gasto com tecnologia está associado aos

resultados comerciais. Assim, as despesas com tecnologia são vistas não como custos, e sim como investimentos. Devido às diferenças fundamentais entre a nuvem (como a taxa de alteração no uso, definição de preço com pagamento conforme o uso, definição de preço em camadas, modelos de definição de preço e informações detalhadas de faturamento e uso) em comparação à operação no local, é essencial que a organização financeira entenda como o uso da nuvem pode afetar aspectos empresariais, incluindo processos de aquisição, rastreamento de incentivos, alocação de custos e demonstrações financeiras.

- Líderes de tecnologia: os líderes de tecnologia (incluindo proprietários de produtos e aplicativos) devem estar cientes dos requisitos financeiros (por exemplo, restrições orçamentárias), bem como dos requisitos de negócios (por exemplo, contratos de nível de serviço). Isso permite que a carga de trabalho seja implementado para atingir os objetivos desejados da organização.

A parceria entre finanças e tecnologia oferece os seguintes benefícios:

- As equipes de finanças e tecnologia têm visibilidade praticamente em tempo real dos custos e do uso.
- As equipes de finanças e tecnologia estabelecem um procedimento operacional padrão para lidar com a variação de gastos na nuvem.
- As partes interessadas nas finanças atuam como consultores estratégicos com relação à forma como o capital é usado para comprar descontos de compromissos (por exemplo, instâncias reservadas ou Savings Plans da AWS) e como a nuvem é usada para expandir a organização.
- Contas a pagar e processos de aquisição existentes são usados com a nuvem.
- As equipes de finanças e tecnologia colaboram na previsão de custos e uso futuros da AWS para alinhar e criar orçamentos organizacionais.
- Melhor comunicação entre organizações por meio de uma linguagem compartilhada e entendimento comum dos conceitos financeiros.

As partes interessadas adicionais dentro da sua organização que devem ser envolvidas em discussões de custo e uso incluem:

- Proprietários de unidades de negócios: os proprietários de unidades de negócios devem compreender o modelo de negócios de nuvem para que possam fornecer orientações tanto para as unidades de negócios quanto para toda a empresa. Esse conhecimento de nuvem é essencial quando há necessidade de prever o crescimento e o uso da carga de trabalho, e ao avaliar opções de compra de longo prazo, como instâncias reservadas ou Savings Plans.

- Equipe de engenharia: uma parceria entre as equipes financeira e de tecnologia é essencial para o desenvolvimento de uma cultura de consciência dos custos que encoraja os engenheiros a agirem em relação ao gerenciamento financeiro na nuvem (CFM). Um dos problemas comuns dos profissionais de CFM ou operações financeiras e das equipes financeiras é fazer com que os engenheiros entendam todos os negócios na nuvem, sigam as práticas recomendadas e tomem as medidas recomendadas.
- Terceiros: se sua organização usa terceiros (por exemplo, consultores ou ferramentas), certifique-se de que eles estejam alinhados com seus objetivos financeiros e possam demonstrar o alinhamento por meio de seus modelos de engajamento e um retorno sobre o investimento (ROI). Terceiros normalmente contribuirão para o relatório e a análise de qualquer carga de trabalho que gerenciem e fornecerão análise de custo de qualquer carga de trabalho que projetem.

Implementar o CFM e obter sucesso requer a colaboração das equipes financeira, comercial e de tecnologia, além de uma mudança na forma como os gastos com nuvem são comunicados e avaliados em toda a organização. Inclua as equipes de engenharia para que façam parte dessas conversas sobre custos e uso em todos os estágios, incentivando-as a seguir as práticas recomendadas e tomar medidas previamente acordadas conforme for apropriado.

Etapas da implementação

- Defina os membros principais: Verifique se todos os membros relevantes de suas equipes de finanças e tecnologia participam da parceria. Os membros financeiros relevantes serão aqueles que interagem com a conta da nuvem. Normalmente serão CFOs, controladores financeiros, planejadores financeiros, analistas de negócios, compras e sourcing. Normalmente, os membros de tecnologia serão proprietários de produtos e aplicativos, gerentes técnicos e representantes de todas as equipes que criam na nuvem. Outros membros podem incluir proprietários de unidades de negócios, como marketing que influenciará o uso de produtos, e terceiros, como consultores para alcançar o alinhamento com seus objetivos e mecanismos e para auxiliar na geração de relatórios.
- Definir tópicos para discussão: Defina os tópicos que são comuns entre as equipes ou que precisarão de um entendimento compartilhado. Siga o custo a partir do momento em que ele é criado, até que a fatura seja paga. Observe todos os membros envolvidos e os processos organizacionais que devem ser aplicados. Compreenda cada etapa ou processo que ele passa e as informações associadas, como modelos de definição de preço disponíveis, definição de preço em camadas, modelos de desconto, orçamento e requisitos financeiros.

- Estabelecer um ritmo regular: Para criar uma parceria financeira e tecnológica, estabeleça uma comunicação regular para criar e manter o alinhamento. O grupo precisa se reunir regularmente para comparar objetivos e métricas. Um ritmo típico envolve analisar o estado da organização, todos os programas em execução no momento e as métricas financeiras e de otimização gerais. Em seguida, as principais workloads são relatadas em mais detalhes.

Recursos

Documentos relacionados:

- Blog de novidades da AWS

COST01-BP03 Estabeleça orçamentos e previsões 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 orientadores de negócios, ou uma combinação deles.

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

Orientação de implementação

os clientes usam a nuvem para obter eficiência, velocidade e agilidade, o que cria uma quantidade altamente variável de custo e utilização. Os custos podem diminuir com o aumento na eficiência da carga de trabalho ou à medida que novas cargas de trabalho e recursos são implantados. É possível ver um aumento no custo quando a eficiência da workload também aumenta, ou quando novas workloads e recursos são implantados. Ou as cargas de trabalho serão escaladas para atender a mais clientes, o que aumenta a utilização e os custos da nuvem. Os recursos estão mais acessíveis do que nunca. A elasticidade da nuvem também traz elasticidade para os custos e as previsões. Os processos de orçamento organizacional existentes devem ser modificados para incorporar essa variabilidade.

Ajuste os processos de previsão e orçamento existentes para se tornarem mais dinâmicos usando um algoritmo baseado em tendências (usando custos históricos como entradas). Você também pode usar algoritmos baseados em orientadores de negócios (por exemplo, lançamentos de novos produtos ou expansão regional) ou uma combinação de tendências e orientadores de negócios.

Uso AWS Budgets para ver orçamentos personalizados no nível granular especificando o período, a recorrência ou a quantidade (fixa ou variável) e adicionando filtros, como serviço, região da AWS e tags. Para manter-se informado sobre a performance de orçamentos existentes, você pode criar e programar o envio regular de relatórios do AWS Budgets para você e para as partes interessadas. Você também pode criar alertas do AWS Budgets com base nos custos reais, cuja natureza é reativa, ou com base nos custos previstos, o que oferece tempo para implementar mitigações de possíveis excessos de custos. Você receberá um alerta quando exceder o custo ou uso, ou se houver previsão de que exceda a quantia orçada.

A AWS oferece a flexibilidade para que você desenvolva processos dinâmicos de previsão e definição orçamentária que vão manter você informado sobre a situação dos custos no que diz respeito a cumprir ou exceder os limites orçamentários.

Uso AWS Cost Explorer para prever os custos em um período futuro definido com base em seu gasto no passado. O mecanismo de previsão do AWS Cost Explorer segmenta seus dados históricos com base em tipos de cobrança (por exemplo, instâncias reservadas) e usa uma combinação de machine learning e modelos baseados em regras para prever os gastos individualmente para todos os tipos de cobrança. Uso AWS Cost Explorer para prever custos de nuvem diários (até três meses) ou mensais (até 12 meses) com base em algoritmos de machine learning aplicados aos seus custos históricos (com base em tendências).

Depois de determinar suas previsões baseadas em tendências usando o Cost Explorer, use o AWS Pricing Calculator para estimar os custos do seu caso de uso da AWS e os custos futuros com base no uso esperado (tráfego, solicitações por segundo, instância do Amazon Elastic Compute Cloud (Amazon EC2) necessária e assim por diante). Você também pode usá-lo para planejar seus gastos, identificar oportunidades de economia e tomar decisões informadas ao usar a AWS.

Uso AWS Cost Anomaly Detection para impedir ou reduzir custos-surpresa e melhorar o controle sem atrasar a inovação. O AWS Cost Anomaly Detection utiliza tecnologias avançadas de machine learning para identificar gastos anormais e causas raiz, para que você tome medidas com rapidez. Com três etapas simplesvocê pode criar seu próprio monitor contextualizado e receber alertas sempre que qualquer gasto anormal for detectado. Deixe os desenvolvedores desenvolverem e permita que o AWS Cost Anomaly Detection monitore seus gastos e reduza o risco de surpresas no faturamento.

Como mencionado na subseção Parceria financeira e tecnológica do pilar de otimização de custos do Well-Architected , é importante ter uma parceria e ritmo entre TI, departamento financeiro e outras partes interessadas para garantir que todos usem as mesmas ferramentas e processos para manter

a consistência. Nas situações em que os orçamentos precisem sofrer alterações, aumentar o ritmo dos pontos de contato pode ajudar na hora de reagir a essas mudanças com mais rapidez.

Etapas da implementação

- Atualizar os processos de previsão e orçamento existentes: implemente algoritmos baseados em tendências ou em orientadores de negócios ou uma combinação de ambos em seus processos de previsão e orçamento.
- Configurar alertas e notificações: Use alertas do AWS Budgets e o Cost Anomaly Detection.
- Realizar revisões regulares com partes interessadas importantes: por exemplo, partes interessadas nos departamentos de TI, financeiro e plataforma, bem como de outras áreas da empresa, para que se alinhem às mudanças no rumo dos negócios e no uso.

Recursos

Documentos relacionados:

- AWS Cost Explorer
- AWS Budgets
- AWS Pricing Calculator
- AWS Cost Anomaly Detection
- AWS License Manager

Exemplos relacionados:

- Lançamento: a previsão baseada em uso já está disponível no AWS Cost Explorer
- Laboratórios do AWS Well-Architected: Governança de custo e uso

COST01-BP04 Implemente o reconhecimento de custos em seus processos organizacionais

Implemente o reconhecimento de custos, crie transparência e contabilize os 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.

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

Orientação de implementação

O reconhecimento de custos deve ser implementado em processos organizacionais novos e existentes. É um dos recursos fundamentais para outras práticas recomendadas. Recomendamos reutilizar e modificar processos existentes sempre que possível, o que minimiza o impacto na agilidade e velocidade. Informe os custos da nuvem para as equipes de tecnologia e os responsáveis por decisões nas equipes financeira e comercial para conscientizar sobre os custos, e estabeleça indicadores-chave de desempenho (KPIs) para as partes interessadas dos departamentos financeiro e comercial. As recomendações a seguir ajudarão a implementar o reconhecimento de custos em sua carga de trabalho:

- Verifique se o gerenciamento de mudanças inclui uma medição de custo para quantificar o impacto financeiro das mudanças. Isso ajuda a abordar de forma proativa as preocupações relacionadas a custos e a destacar as economias de custos.
- Verifique se a otimização de custos é um componente essencial de seus recursos operacionais. Por exemplo, você pode aproveitar os processos existentes de gerenciamento de incidentes para investigar e identificar causas raiz das anomalias de custo e uso ou excessos de custo.
- Acelere a economia de custos e a obtenção de valor empresarial por meio da automação ou de ferramentas. Ao pensar sobre o custo da implementação, enquadre a conversa para incluir um componente de retorno sobre o investimento (ROI) para justificar o investimento de tempo ou dinheiro.
- Aloque os custos de nuvem implementando showbacks ou chargebacks de gastos na nuvem, incluindo gastos com opções de compra baseadas em compromissos, serviços compartilhados e compras de marketplace para impulsionar um consumo da nuvem mais consciente sobre custos.
- Estenda os programas de treinamento e desenvolvimento existentes para incluir treinamento com reconhecimento de custos em toda a organização. Recomendamos que isso inclua treinamento e certificação contínuos. Isso criará uma organização capaz de autogerenciar custos e uso.
- Aproveite ferramentas nativas e gratuitas da AWS, como AWS Cost Anomaly Detection, AWS Budgetse aos relatórios do AWS Budgets.

Quando as organizações adotam sistematicamente práticas de Gerenciamento financeiro na nuvem (CFM), esses comportamentos passam a estar enraizados no modo de trabalho e tomada de decisão. O resultado é uma cultura mais consciente em relação aos custos, desde os desenvolvedores que arquitetam uma nova aplicação concebida na nuvem até gerentes financeiros que analisam o ROI desses novos investimentos na nuvem.

Etapas da implementação

- Identificar processos organizacionais relevantes: Cada unidade organizacional analisa os processos que possui e identifica aqueles que afetam o custo e o uso. Todos os processos que resultam na criação ou no encerramento de um recurso precisam ser incluídos para análise. Procure processos que possam sustentar o reconhecimento de custos na empresa, como gerenciamento de incidentes e treinamento.
- Estabeleça uma cultura com reconhecimento de custos autossustentável. Garanta que todas as partes interessadas relevantes se alinhem ao motivo da mudança e impacto como custo para que entendam os custos da nuvem. Isso vai possibilitar que sua organização estabeleça uma cultura de inovação autossustentável com reconhecimento de custos.
- Atualizar processos com reconhecimento de custos: Cada processo é modificado para ter reconhecimento de custos. O processo pode exigir pré-verificações adicionais, como avaliação do impacto do custo, ou pós-verificações que validam se as mudanças esperadas no custo e no uso ocorreram. Processos de suporte, como treinamento e gerenciamento de incidentes, podem ser estendidos para incluir itens de custo e uso.

Para obter ajuda, fale com especialistas em CFM por meio de sua equipe de conta, ou explore os recursos e os documentos relacionados abaixo.

Recursos

Documentos relacionados:

- Gerenciamento financeiro na nuvem da AWS

Exemplos relacionados:

- Estratégia para um gerenciamento eficiente dos custos da nuvem
- Série de blogs sobre controle de custos n.º 3: Como lidar com o impacto dos custos
- Um guia de introdução ao AWS Cost Management

COST01-BP05 Relatar e notificar sobre a otimização de custos

Configure o AWS Budgets e o AWS Cost Anomaly Detection para fornecer notificações sobre custos e usos em relação às metas. Realize reuniões regulares para analisar o custo-benefício da workload e promover a cultura que reconhece os custos.

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

Orientação de implementação

você deve informar regularmente sobre a otimização de custos e usos dentro da sua organização. Você pode implementar sessões dedicadas para otimização de custos ou incluir otimização de custos em seus ciclos regulares de relatórios operacionais para suas cargas de trabalho. Use serviços e ferramentas para identificar e implementar oportunidades de economia de custo. AWS Cost Explorer fornece painéis e relatórios. Você pode acompanhar seu progresso do custo e do uso em relação a orçamentos configurados com os relatórios do AWS Budgets.

Uso AWS Budgets para definir orçamentos personalizados e rastrear seus custos e uso, para que possa responder rapidamente a alertas recebidos via e-mail ou notificações do Amazon Simple Notification Service (Amazon SNS) se o limite for excedido. Defina seu período de orçamento preferencial como diário, mensal, trimestral ou anual, e crie limites específicos para se manter informado sobre o progresso do uso e dos custos reais e previstos rumo ao limite do orçamento. Você também pode definir de emergência e ações em resposta a esses alertas para que sejam executados automaticamente, ou por meio de um processo de aprovação quando uma meta de orçamento é excedida.

Implemente notificações sobre custo e uso para garantir que alterações no custo e no uso possam ser respondidas rapidamente caso não sejam esperadas. AWS Cost Anomaly Detection permite que você reduza os custos-surpresa e aumente o controle sem desacelerar a inovação. O AWS Cost Anomaly Detection identifica gastos anormais e causas raiz, o que ajuda a reduzir o risco de surpresas no faturamento. Com três etapas simples você pode criar seu próprio monitor contextualizado e receber alertas sempre que qualquer gasto anormal for detectado.

Você também pode usar o Amazon QuickSight com dados do AWS Cost and Usage Report (CUR) para fornecer relatórios altamente personalizados com dados mais granulares. O Amazon QuickSight permite que você programe relatórios e receba e-mails periódicos sobre o relatório de custos com o histórico de custos e uso, ou oportunidades de economia de custo.

Uso AWS Trusted Advisor, que oferece orientação para verificar se os recursos provisionados se alinham com as práticas recomendadas da AWS para otimização de custo.

Crie periodicamente relatórios que contêm um destaque de Savings Plans, instâncias reservadas e recomendações de dimensionamento do Amazon Elastic Compute Cloud (Amazon EC2) do AWS Cost Explorer para começar a reduzir o custo associado a workloads estacionárias e recursos ociosos ou subutilizados. Identifique e recupere os gastos associados ao desperdício de

recursos implantados na nuvem. O desperdício na nuvem ocorre quando recursos dimensionados incorretamente são criados ou quando se observa padrões de uso diferentes do esperado. Siga as práticas recomendadas da AWS para reduzir o desperdício e otimizar e economizar os custos da nuvem.

Gere relatórios regularmente para melhorar as opções de compra de recursos a fim de reduzir os custos unitários das workloads. Opções de compra como Savings Plans, instâncias reservadas ou instâncias spot do Amazon EC2 oferecem as maiores economias para workloads tolerantes a falhas e permitem que as partes interessadas (proprietários de negócios e equipes financeiras e de tecnologia) façam parte das conversas sobre comprometimento.

Compartilhe os relatórios que contêm oportunidades ou anúncios de novos lançamentos que possam ajudar você a reduzir o custo total de propriedade (TCO) da nuvem. Adote novos serviços, regiões, recursos, soluções ou maneiras de obter mais reduções de custo.

Etapas da implementação

- Configurar o AWS Budgets: configure o AWS Budgets em todas as contas para a sua workload. Defina um orçamento para o gasto total da conta e outro para a carga de trabalho usando tags.
  - Laboratórios do Well-Architected: Governança de custo e uso
- Relatório sobre otimização de custos: Configure um ciclo regular para discutir e analisar a eficiência da carga de trabalho. Usando as métricas estabelecidas, informe sobre as métricas obtidas e o custo de alcançá-las. Identifique e corrija quaisquer tendências negativas e identifique tendências positivas que podem ser promovidas em toda a organização. Os relatórios devem envolver representantes das equipes de aplicativos e dos proprietários, das finanças e da gerência.
  - Laboratórios do Well-Architected: Visualização

Recursos

Documentos relacionados:

- AWS Cost Explorer
- AWS Trusted Advisor
- AWS Budgets
- Práticas recomendadas do AWS Budgets
- Amazon CloudWatch

- AWS CloudTrail
- Análises do Amazon S3
- AWS Cost and Usage Report

Exemplos relacionados:

- Laboratórios do Well-Architected: Governança de custo e uso
- Laboratórios do Well-Architected: Visualização
- Principais formas de começar a otimizar seus custos de nuvem da AWS

COST01-BP06 Monitore custos proativamente

Implemente ferramentas e painéis para monitorar os custos proativamente para a carga de trabalho. Analise regularmente os custos com ferramentas configuradas ou prontas para usar em vez de apenas analisar os custos e as categorias quando receber notificações. O monitoramento e a análise proativa dos custos ajuda a identificar tendências positivas e permite que você as promova em toda a organização.

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

Orientação de implementação

é recomendável monitorar custos e uso proativamente em sua organização, e não apenas quando há exceções ou anomalias. Painéis altamente visíveis em todo o escritório ou ambiente de trabalho garantem que as principais pessoas tenham acesso às informações necessárias e indicam o foco da organização na otimização de custos. Os painéis visíveis permitem promover ativamente resultados bem-sucedidos e implementá-los em toda a organização.

Crie uma rotina diária ou frequente de uso do AWS Cost Explorer ou de qualquer outro painel, como o Amazon QuickSight , para ver os custos e analisar de forma proativa. Analise o uso e os custos dos serviços da AWS na conta da AWS, no nível da workload ou em um serviço específico da AWS com agrupamento e filtragem, e valide se estão dentro do esperado ou não. Use a granularidade no nível de hora e recurso e as tags para filtrar e identificar os custos incorridos para os principais recursos. Você também pode criar seus próprios relatórios com o painel de inteligência de custos, uma solução do Amazon QuickSight desenvolvida por arquitetos de soluções da AWS, e comparar os orçamentos com o uso e os custos reais.

Etapas da implementação

- Relatório sobre otimização de custos: Configure um ciclo regular para discutir e analisar a eficiência da carga de trabalho. Usando as métricas estabelecidas, informe sobre as métricas obtidas e o custo de alcançá-las. Identifique e corrija quaisquer tendências negativas e identifique tendências positivas a serem promovidas em toda a organização. Os relatórios devem envolver representantes das equipes de aplicativos e dos proprietários, das finanças e da gerência.
- Crie e habilite a granularidade diária do AWS Budgets para o uso e os custos a fim de tomar medidas oportunas para impedir quaisquer possíveis excessos de custo: o AWS Budgets permite que você configure notificações de alerta, para que permaneça informado se qualquer tipo de orçamento sair dos limites pré-configurados. A melhor forma de aproveitar o AWS Budgets é definir o custo e o uso esperados como limites, para que qualquer coisa acima do seu orçamento seja considerada excesso.
- Crie AWS Cost Anomaly Detection para o monitor de custos: AWS Cost Anomaly Detection usa tecnologia avançada de machine learning para identificar gastos anormais e causas raiz, para que você possa agir rapidamente. Permite que você configure monitores de custo que definem os segmentos de gastos que deseja avaliar (por exemplo, serviços individuais da AWS, contas de membros, tags de alocação de custo e categorias de custo) e permite que você defina quando, onde e como recebe notificações de alerta. Para cada monitor, anexe várias assinaturas de alertas para proprietários de negócios e equipes de tecnologia, incluindo um nome, um limite de impacto do custo e a frequência de alerta (alertas individuais, resumo diário, resumo semanal) para cada assinatura.
- Use o AWS Cost Explorer ou integre seus dados do AWS Cost and Usage Report (CUR) com painéis do Amazon QuickSight para visualizar os custos da organização: o AWS Cost Explorer conta com uma interface fácil de usar que permite que você visualize, entenda e gerencie os custos e o uso da AWS com o passar do tempo. O painel de inteligência de custos é um painel personalizável e acessível para ajudar a criar a base de sua própria ferramenta de gerenciamento e otimização dos custos.

Recursos

Documentos relacionados:

- AWS Budgets
- AWS Cost Explorer
- Orçamentos diários para custos e uso
- AWS Cost Anomaly Detection

Exemplos relacionados:

- Laboratórios do Well-Architected: Visualização
- Laboratórios do Well-Architected: Visualização avançada
- Laboratórios do Well-Architected: Painéis de inteligência de nuvem
- Laboratórios do Well-Architected: Visualização de custos
- Alerta do AWS Cost Anomaly Detection com Slack

COST01-BP07 Manter-se atualizado com os novos lançamentos de serviços

Consulte regularmente especialistas ou parceiros da AWS para considerar quais serviços e recursos oferecem menor custo. Analise os blogs da AWS e outras fontes de informação.

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

Orientação de implementação

A AWS adiciona novos recursos constantemente para que você possa aproveitar as tecnologias mais recentes a fim de experimentar e inovar com maior rapidez. Você pode implementar novos serviços e recursos da AWS para aumentar a eficiência de custos na workload. Confira regularmente o Gerenciamento de custos da AWS, o Blog de novidades da AWS, o Blog de gerenciamento de custos da AWSe aos Novidades da AWS para obter informações sobre novos lançamentos de serviços e recursos. As postagens de Novidades oferecem uma breve visão geral de todos os anúncios de serviços, recursos e expansões de regiões da AWS à medida que são lançados.

Etapas da implementação

- Inscrever-se em blogs: Acesse as páginas de blogs da AWS e inscreva-se em Novidades e em outros blogs relevantes. Você pode inscrever-se na página de preferências de comunicação com seu endereço de e-mail.
- Inscrever-se para receber as Novidades da AWS: Confira regularmente o Blog de novidades da AWS e Novidades da AWS para obter informações sobre novos lançamentos de serviços e recursos. Assine o feed RSS, ou use seu e-mail para ficar por dentro dos anúncios e lançamentos.
- Seguir as reduções de preço da AWS: cortes regulares nos preços de todos os nossos serviços são uma prática padrão que a AWS usa para passar os benefícios econômicos obtidos pela nossa escala aos clientes. Até abril de 2022, a AWS já reduziu preços 115 vezes desde seu lançamento em 2006. Se você tiver qualquer decisão comercial pendente por motivos de preço, poderá

reavaliar depois de reduções de preços e novas integrações de serviços. Você pode saber mais sobre nossos esforços anteriores para redução de preços, incluindo instâncias do Amazon Elastic Compute Cloud (Amazon EC2), na categoria de redução de preços do Blog de novidades da AWS.

- Eventos e reuniões da AWS: participe da conferência local da AWS e de qualquer reunião local com outras organizações da área. Se não puder participar presencialmente, tente participar dos eventos virtuais para ouvir mais de especialistas da AWS e casos de negócios de outros clientes.
- Reunir-se com a equipe da sua conta: programe um ritmo regular com a equipe de contas, encontre-se com ela e discuta as tendências do setor e os serviços da AWS. Fale com o gerente de contas, o arquiteto de soluções e a equipe de suporte.

Recursos

Documentos relacionados:

- Gerenciamento de custos da AWS
- Novidades da AWS
- Blog de novidades da AWS

Exemplos relacionados:

- Amazon EC2: 15 anos de otimização e economia de custos de TI
- Blog de novidades da AWS: redução de preços

Reconhecimento de despesas e usos

Perguntas

- COST 2 Como você controla o uso?
- COST 3 Como você monitora o uso e os custos?
- COST 4 Como você desativa recursos?

COST 2 Como você controla 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.

Práticas recomendadas

- COST02-BP01 Desenvolva políticas com base nos requisitos da sua organização
- COST02-BP02 Implemente objetivos e metas
- COST02-BP03 Implemente uma estrutura de conta
- COST02-BP04 Implemente grupos e funções
- COST02-BP05 Implementar controles de custos
- COST02-BP06 Acompanhe o ciclo de vida do projeto

COST02-BP01 Desenvolva políticas com base nos requisitos da sua 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.

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

Orientação de implementação

Entender os custos e os orientadores da sua organização é essencial para gerenciar seus custos e uso com eficácia e identificar oportunidades de redução de custos. Normalmente, as organizações operam várias cargas de trabalho executadas por várias equipes. Essas equipes podem estar em diferentes unidades da organização, cada uma com o próprio fluxo de receita. A capacidade de atribuir custos de recursos a workloads, à organização individual ou aos proprietários do produto gera um comportamento eficiente do uso e ajuda a reduzir o desperdício. O monitoramento preciso de custos e uso permite que você entenda como as unidades e os produtos da organização são rentáveis e permite que você tome decisões mais embasadas sobre onde alocar recursos dentro da sua organização. A conscientização sobre o uso em todos os níveis da organização é essencial para promover mudanças, pois a mudança no uso gera mudanças no custo. Considere adotar uma abordagem multifacetada para se tornar ciente do seu uso e das suas despesas.

o primeiro passo para executar governança é usar os requisitos da sua organização para desenvolver políticas para o uso da nuvem. Essas políticas definem como sua organização usa a nuvem e como os recursos são gerenciados. As políticas devem cobrir todos os aspectos de recursos e cargas de trabalho relacionados ao custo ou uso, incluindo criação, modificação e desativação durante a vida útil do recurso.

As políticas devem ser simples, para que sejam facilmente compreendidas e possam ser implementadas com eficácia em toda a organização. Comece com políticas amplas e de alto nível,

como em qual região geográfica o uso é permitido ou horários do dia em que os recursos devem estar em execução. Refine gradualmente as políticas para as várias unidades organizacionais e cargas de trabalho. As políticas comuns incluem quais serviços e recursos podem ser usados (por exemplo, armazenamento de menor performance em ambientes de teste ou desenvolvimento) e quais tipos de recursos podem ser usados por diferentes grupos (por exemplo, o maior tamanho de recursos em uma conta de desenvolvimento é médio).

Etapas da implementação

- Reuna-se com membros da equipe: Para desenvolver políticas, faça com que todos os membros da equipe da sua organização especifiquem seus requisitos e os documentem de acordo. Adote uma abordagem iterativa iniciando uma refinação ampla e contínua para as menores unidades em cada etapa. Os membros da equipe incluem aqueles com interesse direto na carga de trabalho, como unidades da organização ou proprietários de aplicativos, bem como grupos de apoio, como equipes de segurança e finanças.
- Defina locais para sua workload: Defina onde sua carga de trabalho opera, incluindo o país e a área dentro do país. Essas informações são usadas para mapear zonas de disponibilidade e Regiões da AWS.
- Defina e agrupe serviços e recursos: Defina os serviços que as cargas de trabalho exigem. Para cada serviço, especifique os tipos, o tamanho e o número de recursos necessários. Defina grupos para os recursos por função, como servidores de aplicativos ou armazenamento de banco de dados. Os recursos podem pertencer a vários grupos.
- Defina e agrupe os usuários por função: Defina os usuários que interagem com a carga de trabalho, concentrando-se no que eles fazem e em como usam a carga de trabalho, não em quem são ou na posição deles na organização. Agrupe usuários ou funções semelhantes. Você pode usar as políticas gerenciadas da AWS como um guia.
- Defina as ações: Usando os locais, recursos e usuários identificados anteriormente, defina as ações que são exigidas por cada um para alcançar os resultados da carga de trabalho ao longo do tempo de vida (desenvolvimento, operação e desativação). Identifique as ações com base nos grupos, e não nos elementos individuais nos grupos, em cada local. Comece amplamente com leitura ou gravação e, em seguida, refine ações específicas para cada serviço.
- Defina o período de análise: As cargas de trabalho e os requisitos organizacionais podem mudar ao longo do tempo. Defina a programação de análise da carga de trabalho para garantir que ela permaneça alinhada com as prioridades organizacionais.

- Documente as políticas: Verifique se as políticas que foram definidas estão acessíveis conforme exigido pela sua organização. Essas políticas são usadas para implementar, manter e auditar o acesso de seus ambientes.

Recursos

Documentos relacionados:

- Políticas gerenciadas da AWS para funções de trabalho
- Estratégia de faturamento de várias contas da AWS
- Ações, recursos e chaves de condição para serviços da AWS
- Produtos da nuvem
- Controle o acesso a Regiões da AWS usando políticas da IAM
- Regiões e AZs de infraestruturas globais

COST02-BP02 Implemente 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.

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

Orientação de implementação

desenvolva objetivos e metas de custo e uso para a sua organização. Os objetivos fornecem orientações e direcionamento para a sua organização sobre os resultados esperados. As metas fornecem resultados mensuráveis específicos a serem alcançados. Um exemplo de um objetivo é: o uso da plataforma deve aumentar significativamente, com apenas um pequeno aumento (não linear) no custo. Um exemplo de meta é: um aumento de 20% no uso da plataforma, com um aumento de menos de 5% nos custos. Outro objetivo comum é que as cargas de trabalho precisam ser mais eficientes a cada seis meses. A meta acompanhante seria que o custo por saída da carga de trabalho precisa diminuir em 5% a cada 6 meses.

Um objetivo comum para cargas de trabalho na nuvem é aumentar a eficiência da carga de trabalho, que diminuirá o custo por resultado comercial da carga de trabalho ao longo do tempo. É recomendável implementar essa meta para todas as workloads e também definir uma meta como

um aumento de 5% na eficiência a cada 6 a 12 meses. Isso pode ser obtido na nuvem por meio da criação de recursos na otimização de custos e do lançamento de novos serviços e recursos de serviços.

Etapas da implementação

- Defina níveis de uso esperados: Concentre-se nos níveis de uso para começar. Envolva-se com proprietários de aplicações, marketing e equipes de negócios maiores para entender quais serão os níveis de uso esperados para a workload. Como a demanda do cliente mudará ao longo do tempo, e se haverá alterações devido a aumentos sazonais ou campanhas de marketing.
- Defina custos e recursos de workload: Com os níveis de uso definidos, quantifique as alterações nos recursos da carga de trabalho necessárias para atender a esses níveis de uso. Pode ser necessário aumentar o tamanho ou o número de recursos para um componente de carga de trabalho, aumentar a transferência de dados ou alterar componentes de carga de trabalho para um serviço diferente em um nível específico. Especifique quais serão os custos em cada um desses pontos principais e quais serão as alterações no custo quando houver alterações no uso.
- Defina metas empresariais: Combine o resultado das mudanças esperadas no uso e no custo com as mudanças esperadas na tecnologia ou qualquer programa que você esteja executando e desenvolva metas para a carga de trabalho. As metas devem abordar o uso, o custo e a relação entre os dois. Verifique se há programas organizacionais, por exemplo, criação de recursos como treinamento e educação, se houver alterações esperadas no custo sem alterações no uso.
- Defina objetivos: Para cada uma das metas definidas, especifique um objetivo mensurável. Se uma meta for aumentar a eficiência na workload, o objetivo quantificará a quantidade da melhoria, típica nos resultados de negócios por cada dólar gasto, e quando ela será entregue.

Recursos

Documentos relacionados:

- Políticas gerenciadas da AWS para funções de trabalho
- Estratégia de faturamento de várias contas da AWS
- Controle o acesso a Regiões da AWS usando políticas da IAM

COST02-BP03 Implemente 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.

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

Orientação de implementação

A AWS tem uma estrutura de conta que segue o modelo “um pai para muitos filhos” e que é comumente conhecida como conta de gerenciamento (o pai, anteriormente conta pagante) com conta-membro (o filho, anteriormente conta vinculada). Uma prática recomendada é sempre ter pelo menos uma conta de gerenciamento com uma conta-membro, independentemente do tamanho ou uso da sua organização. Todos os recursos de carga de trabalho devem residir somente em contas membro.

Não há uma resposta geral para quantas contas da AWS você deve ter. Avalie seus modelos de custo e operacionais atuais e futuros para garantir que a estrutura de suas contas da AWS reflita os objetivos da sua organização. Algumas empresas criam várias contas da AWS por motivos comerciais, por exemplo:

- O isolamento administrativo e/ou fiscal e de faturamento é necessário entre unidades da organização, centros de custo ou cargas de trabalho específicas.
- Os limites de serviço da AWS são definidos para que sejam específicos para determinadas workloads.
- Há um requisito de isolamento e separação entre cargas de trabalho e recursos.

No AWS Organizations, o faturamento consolidado cria a construção entre uma ou mais contas-membros e a conta de gerenciamento. As contas membro permitem que você isole e diferencie seu custo e uso por grupos. Uma prática comum é ter contas membro separadas para cada unidade da organização (como finanças, marketing e vendas), ou para cada ciclo de vida do ambiente (como desenvolvimento, teste e produção) ou para cada carga de trabalho (carga de trabalho a, b e c) e, em seguida, agregar essas contas vinculadas usando o faturamento consolidado.

O faturamento consolidado permite consolidar o pagamento de várias contas membro da AWS em uma única conta de gerenciamento, sem deixar de oferecer visibilidade para a atividade de cada conta vinculada. Como os custos e o uso são agregados na conta de gerenciamento, isso permite maximizar seus descontos por volume de serviço e maximizar o uso de seus descontos de compromisso (Savings Plans e instâncias reservadas) para obter os descontos mais altos.

AWS Control Tower pode instalar e configurar rapidamente várias contas da AWS, garantindo que a governança esteja alinhada com os requisitos da sua organização.

Etapas da implementação

- Defina requisitos de separação: Os requisitos de separação são uma combinação de vários fatores, que incluem segurança, confiabilidade e construções financeiras. Trabalhe em cada fator em ordem e especifique se a carga de trabalho ou o ambiente dela deve ser separado de outras cargas de trabalho. A segurança garante que os requisitos de acesso e dados sejam cumpridos. A confiabilidade garante que os limites sejam gerenciados para que os ambientes e as cargas de trabalho não afetem outros. Construções financeiras garantem que haja separação e prestação de contas financeiras rigorosas. Exemplos comuns de separação são cargas de trabalho de produção e teste executadas em contas separadas ou o uso de uma conta separada para que os dados da fatura e do faturamento possam ser fornecidos a uma organização terceirizada.
- Defina requisitos de agrupamento: Os requisitos de agrupamento não modificam os requisitos de separação, mas são usados para auxiliar o gerenciamento. Agrupe ambientes semelhantes ou cargas de trabalho que não exigem separação. Um exemplo disso é o agrupamento de vários ambientes de teste ou desenvolvimento de uma ou mais cargas de trabalho.
- Defina a estrutura da conta: Usando essas separações e agrupamentos, especifique uma conta para cada grupo e mantenha os requisitos de separação. Essas contas são suas contas de membros ou vinculadas. Ao agrupar essas contas-membros em uma única conta de gerenciamento ou pagante, você combina o uso, o que permite maiores descontos por volume em todas as contas e fornece uma única fatura para todas as contas. É possível separar dados de faturamento e fornecer a cada conta de membro uma visualização individual dos dados de faturamento. Se uma conta-membro não precisar ter os dados de uso ou faturamento visíveis para nenhuma outra conta, ou se uma fatura separada da AWS for necessária, defina várias contas de gerenciamento ou pagantes. Nesse caso, cada conta-membro tem a própria conta de gerenciamento ou pagante. Os recursos devem sempre ser colocados em contas-membros ou vinculadas. As contas de gerenciamento ou pagantes devem ser usadas somente para gerenciamento.

Recursos

Documentos relacionados:

- Políticas gerenciadas da AWS para funções de trabalho
- Estratégia de faturamento de várias contas da AWS
- Controle o acesso a Regiões da AWS usando políticas da IAM
- AWS Control Tower
- AWS Organizations
- Faturamento consolidado

Exemplos relacionados:

- Dividir o CUR e compartilhar o acesso

COST02-BP04 Implemente 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.

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

Orientação de implementação

Depois de desenvolver políticas, você pode criar funções e grupos lógicos de usuários em sua organização. Isso permite que você atribua permissões e controle o uso. Comece com agrupamentos de pessoas de alto nível. Normalmente isso se alinha a unidades organizacionais e funções de trabalho (por exemplo, administrador de sistemas no departamento de TI ou controlador financeiro). Os grupos juntam pessoas que realizam tarefas semelhantes e precisam de acesso semelhante. As funções definem o que um grupo deve fazer. Por exemplo, um administrador de sistemas em TI requer acesso para criar todos os recursos, mas um membro da equipe de análise só precisa criar recursos de análise.

Etapas da implementação

- Implemente grupos: Usando os grupos de usuários definidos em suas políticas organizacionais, implemente os grupos correspondentes, se necessário. Consulte o pilar Segurança para obter as melhores práticas sobre usuários, grupos e autenticação.
- Implemente funções e políticas: Usando as ações definidas em suas políticas organizacionais, crie as funções e as políticas de acesso necessárias. Consulte o pilar Segurança para obter as melhores práticas sobre funções e políticas.

Recursos

Documentos relacionados:

- Políticas gerenciadas da AWS para funções de trabalho
- Estratégia de faturamento de várias contas da AWS

- Controle o acesso a Regiões da AWS usando políticas da IAM
- Pilar Segurança do AWS Well-Architected

Exemplos relacionados:

- Laboratório do Well-Architected: Identidade e acesso básico

COST02-BP05 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 conforme definido pelos requisitos da organização: por exemplo, controle o acesso a regiões ou tipos de recursos com políticas do AWS Identity and Access Management (IAM).

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

Orientação de implementação

Uma primeira etapa comum na implementação de controles de custo é configurar notificações quando eventos de custo ou uso ocorrerem fora das políticas da organização. Isso permite que você aja rapidamente e verifique se é necessária uma ação corretiva, sem restringir ou afetar negativamente cargas de trabalho ou novas atividades. Depois de conhecer os limites da carga de trabalho e do ambiente, você pode impor a governança. Na AWS, as notificações são realizadas com AWS Budgets, o que permite definir um orçamento mensal para seus custos, uso e descontos de compromisso da AWS (Savings Plans e instâncias reservadas). Você pode criar orçamentos em um nível de custo agregado (por exemplo, todos os custos) ou em um nível mais granular, onde você inclui apenas dimensões específicas, como contas vinculadas, serviços, tags ou zonas de disponibilidade.

Como segunda etapa, você pode aplicar políticas de governança na AWS por meio do AWS Identity and Access Management (IAM) e Políticas de controle de serviço (SCP) do AWS Organizations. O IAM permite gerenciar com segurança o acesso aos serviços e recursos da AWS. Usando o IAM, você pode controlar quem pode criar e gerenciar recursos da AWS, os tipos de recursos que podem ser criados e onde eles podem ser criados. Isso minimiza a criação de recursos que não são necessários. Use as funções e os grupos criados anteriormente e atribua políticas do IAM para impor o uso correto. A SCP oferece controle central sobre o número máximo de permissões disponíveis para todas as contas na sua organização, garantindo que suas contas permaneçam dentro das diretrizes de controle de acesso. As SCPs estão disponíveis somente em uma organização com

todos os recursos habilitados, e você pode configurar as SCPs para negar ou permitir ações para contas membro por padrão. Consulte o whitepaper sobre o Pilar Segurança do Well-Architected para obter mais detalhes sobre a implementação do gerenciamento de acesso.

A governança também pode ser implementada por meio do gerenciamento do Service Quotas. Ao garantir que o Service Quotas seja configurado com o mínimo de sobrecarga e mantido com precisão, você pode minimizar a criação de recursos fora dos requisitos da sua organização. Para conseguir isso, você deve entender a rapidez com que seus requisitos podem mudar, compreender projetos em andamento (criação e desativação de recursos) e considerar a rapidez com que as alterações de cota podem ser implementadas. O Service Quotas pode ser usado para aumentar suas cotas quando necessário.

Etapas da implementação

- Implementar notificações sobre gastos: Usando as políticas da organização definidas, crie orçamentos da AWS para fornecer notificações quando os gastos estiverem fora de suas políticas. Configure vários orçamentos de custos, um para cada conta, que o notifica sobre os gastos gerais da conta. Em seguida, configure orçamentos de custos adicionais dentro de cada conta para unidades menores dentro da conta. Essas unidades variam de acordo com a estrutura da sua conta. Alguns exemplos comuns são Regiões da AWS, workloads (usando tags) ou serviços da AWS. Configure uma lista de distribuição de e-mails como o destinatário das notificações, e não uma conta de e-mail de uma pessoa. Você pode configurar um orçamento real para quando um valor for ultrapassado ou usar um orçamento previsto para notificar sobre o uso previsto.
- Implementar controles de uso: Usando as políticas da organização definidas, implemente políticas e perfis do IAM para especificar quais ações os usuários podem e não podem executar. Várias políticas organizacionais podem ser incluídas em uma política da AWS. Da mesma forma que você definiu políticas, comece amplamente e, em seguida, aplique controles mais granulares em cada etapa. Os limites de serviço também são um controle eficaz do uso. Implemente os limites de serviço corretos em todas as suas contas.

Recursos

Documentos relacionados:

- Políticas gerenciadas da AWS para funções de trabalho
- Estratégia de faturamento de várias contas da AWS
- Controle o acesso a Regiões da AWS usando políticas do IAM

Exemplos relacionados:

- Laboratórios do Well-Architected: Governança de custo e uso
- Laboratórios do Well-Architected: Governança de custo e uso

COST02-BP06 Acompanhe 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.

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

Orientação de implementação

Certifique-se de acompanhar todo o ciclo de vida da carga de trabalho. Isso garante que, quando cargas de trabalho ou componentes de carga de trabalho não forem mais necessários, eles possam ser desativados ou modificados. Isso é especialmente útil quando você lança novos serviços ou recursos. As cargas de trabalho e os componentes existentes podem parecer estar em uso, mas devem ser desativados para redirecionar os clientes para o novo serviço. Observe estágios anteriores das cargas de trabalho - depois que uma carga de trabalho está em produção, os ambientes anteriores podem ser desativados ou significativamente reduzidos na capacidade até que sejam necessários novamente.

A AWS fornece uma série de serviços de gerenciamento e governança que você pode usar para o rastreamento do ciclo de vida da entidade. Você pode usar o AWS Config ou AWS Systems Manager para fornecer um inventário detalhado dos recursos e da configuração da AWS. Recomendamos que você o integre com seus sistemas existentes de gerenciamento de projetos ou ativos para acompanhar projetos e produtos ativos em sua organização. A combinação do seu sistema atual com o conjunto completo de eventos e métricas fornecido pela AWS permite criar uma visão de eventos de ciclo de vida significativos e gerenciar recursos proativamente para reduzir custos desnecessários.

Consulte o whitepaper sobre o Pilar Excelência operacional do Well-Architected para obter mais detalhes sobre a implementação do rastreamento do ciclo de vida de entidades.

Etapas da implementação

- Execute análises de workload: Conforme definido por suas políticas organizacionais, audite seus projetos existentes. A quantidade de esforço utilizado na auditoria deve ser proporcional ao risco aproximado, valor ou custo para a organização. As principais áreas a serem incluídas na

auditoria seriam riscos para a organização de um incidente ou interrupção, valor ou contribuição para a organização (medidos em receita ou reputação da marca), custo da carga de trabalho (medido como custo total de recursos e custos operacionais) e uso da carga de trabalho (medido em número de resultados da organização por unidade de tempo). Se essas áreas mudarem ao longo do ciclo de vida, serão necessários ajustes na carga de trabalho, como desativação total ou parcial.

Recursos

Documentos relacionados:

- AWS Config
- AWS Systems Manager
- Políticas gerenciadas da AWS para funções de trabalho
- Estratégia de faturamento de várias contas da AWS
- Controle o acesso a Regiões da AWS usando políticas da IAM

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.

Práticas recomendadas

- COST03-BP01 Configurar fontes de informações detalhadas
- COST03-BP02 Identificar categorias de atribuição de custos
- COST03-BP03 Estabelecer métricas da organização
- COST03-BP04 Configure as ferramentas de faturamento e gerenciamento de custos
- COST03-BP05 Adicionar informações da organização ao custo e ao uso
- COST03-BP06 Alocar custos com base nas métricas de workload

COST03-BP01 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.

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

Orientação de implementação

Habilite a granularidade por hora no AWS Cost Explorer e crie um AWS Cost and Usage Report (CUR). Essas fontes de dados oferecem a visualização mais precisa do custo e do uso em toda a organização. O CUR fornece granularidade de uso diário ou por hora, taxas, custos e atributos de uso para todos os serviços da AWS cobráveis. Todas as dimensões possíveis estão no CUR, incluindo: marcação, localização, atributos de recurso e IDs de conta.

Configure seu CUR com as seguintes personalizações:

- Incluir IDs de recurso
- Atualizar automaticamente o CUR
- Granularidade por hora
- Versionamento: Substituir relatório existente
- Integração de dados: Amazon Athena (formato Parquet e compactação)

Uso AWS Glue para preparar os dados para análise e use o Amazon Athena para executar a análise de dados, usando SQL para consultar os dados. Você também pode usar o Amazon QuickSight para criar visualizações personalizadas e complexas e distribuí-las em toda a organização.

Etapas da implementação

- Configurar o Relatório de Custos e Uso: Usando o console de faturamento, configure pelo menos um relatório de custos e uso. Configure um relatório com granularidade por hora que inclua todos os identificadores e IDs de recursos. Você também pode criar outros relatórios com diferentes granularidades para fornecer informações resumidas de alto nível.
- Configurar a granularidade por hora no Cost Explorer: Usando o console de faturamento, habilite Por hora e Dados no nível do recurso.

> **Note**
>
> Haverá custos associados à habilitação desse recurso. Consulte a definição de preço para obter mais informações.

- Configurar o registro em log das aplicações: Verifique se a aplicação registra cada resultado empresarial entregue para que possa ser acompanhado e medido. Verifique se a granularidade

desses dados é pelo menos por hora para que corresponda aos dados de custo e uso. Consulte o pilar Excelência operacional do Well-Architected para obter mais detalhes sobre registro em log e monitoramento.

Recursos

Documentos relacionados:

- Configuração de conta da AWS
- AWS Cost and Usage Report (CUR)
- AWS Glue
- Amazon QuickSight
- Definição de preço do Gerenciamento de Custos da AWS
- Marcação de recursos da AWS
- Análise de custos com o AWS Budgets
- Análise de custos com o Cost Explorer
- Gerenciamento do Relatório de Custos e Uso da AWS
- pilar Excelência operacional do Well-Architected

Exemplos relacionados:

- Configuração de conta da AWS

COST03-BP02 Identificar categorias de atribuição de custos

Identifique as categorias de organização que podem ser usadas para alocar custos dentro da organização.

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

Orientação de implementação

trabalhe com sua equipe financeira e outras partes interessadas relevantes para compreender os requisitos de como os custos devem ser alocados em sua organização. Os custos da carga de trabalho devem ser alocados durante todo o ciclo de vida, incluindo desenvolvimento,

teste, produção e desativação. Entenda como os custos incorridos para o aprendizado, o desenvolvimento da equipe e a criação de ideias são atribuídos na organização. Isso pode ser útil para alocar corretamente contas usadas para essa finalidade para orçamentos de treinamento e desenvolvimento, em vez de orçamentos genéricos de custo de TI.

Etapas da implementação

- Definir as categorias da sua organização: Conheça as partes interessadas para definir categorias que reflitam a estrutura e os requisitos da sua organização. Eles serão mapeados diretamente para a estrutura das categorias financeiras existentes, como unidade de negócios, orçamento, centro de custo ou departamento. Veja os resultados que a nuvem oferece para a sua empresa, como treinamento ou educação, já que também são categorias de organização. Várias categorias podem ser atribuídas a um recurso, e um recurso pode estar em várias categorias diferentes. Portanto, defina quantas categorias forem necessárias.
- Definir suas categorias funcionais: Conheça as partes interessadas para definir categorias que reflitam as funções que você tem dentro da sua empresa. Podem ser os nomes da carga de trabalho ou do aplicativo e o tipo de ambiente, como produção, teste ou desenvolvimento. Várias categorias podem ser atribuídas a um recurso, e um recurso pode estar em várias categorias diferentes. Portanto, defina quantas categorias forem necessárias.

Recursos

Documentos relacionados:

- Marcação de recursos da AWS
- Análise de custos com o AWS Budgets
- Análise de custos com o Cost Explorer
- Gerenciamento do Relatório de Custos e Uso da AWS

COST03-BP03 Estabelecer métricas da organização

Estabeleça as métricas da organização que são necessárias para esta carga de trabalho. Exemplo de métricas de uma workload são relatórios de clientes produzidos ou páginas da Web veiculadas aos clientes.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

Entenda como a saída da carga de trabalho é medida em relação ao sucesso empresarial. Cada carga de trabalho normalmente tem um pequeno conjunto de saídas principais que indicam performance. Se você tiver uma carga de trabalho complexa com muitos componentes, poderá priorizar a lista ou definir e rastrear métricas para cada componente. Trabalhe com suas equipes para entender quais métricas usar. Essa unidade será usada para compreender a eficiência da carga de trabalho ou o custo de cada saída de negócios.

Etapas da implementação

- Definir resultados da workload: Reúna-se com as partes interessadas da empresa e defina os resultados para a carga de trabalho. Essas são medidas principais de uso do cliente e devem ser métricas de negócios, e não técnicas. Deve haver um pequeno número de métricas de alto nível (menos de cinco) por carga de trabalho. Se a carga de trabalho produzir vários resultados para diferentes casos de uso, agrupe-os em uma única métrica.
- Definir os resultados do componente da workload: Opcionalmente, se você tiver uma carga de trabalho grande e complexa ou puder facilmente dividir sua carga de trabalho em componentes (como microsserviços) com entradas e saídas bem definidas, defina métricas para cada componente. O esforço deve refletir o valor e o custo do componente. Comece com os maiores componentes e trabalhe em direção aos componentes menores.

Recursos

Documentos relacionados:

- Marcação de recursos da AWS
- Análise de custos com o AWS Budgets
- Análise de custos com o Cost Explorer
- Gerenciamento do Relatório de Custos e Uso da AWS

COST03-BP04 Configure as ferramentas de faturamento e gerenciamento de custos

Configure o AWS Cost Explorer e o AWS Budgets de acordo com as políticas da organização.

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

Orientação de implementação

Para modificar o uso e ajustar os custos, cada pessoa em sua organização deve ter acesso às suas informações de custo e uso. Recomendamos que todas as cargas de trabalho e equipes tenham as seguintes ferramentas configuradas ao usar a nuvem:

- Relatórios: resumo de todas as informações de custo e uso
- Notificações: forneça notificações quando o custo ou o uso estiverem fora dos limites definidos.
- Estado atual: configure um painel mostrando os níveis atuais de custo e uso. O painel deve estar disponível em um local altamente visível dentro do ambiente de trabalho (semelhante a um painel de operações).
- Tendências: fornecem o recurso para mostrar a variabilidade de custo e uso ao longo do período de tempo necessário, com a granularidade necessária.
- Previsões: fornecem o recurso para mostrar custos futuros estimados.
- Rastreamento: mostra o custo e o uso atuais em relação a metas ou objetivos configurados.
- Análises: fornecem a capacidade para os membros da equipe fazerem análises personalizadas e detalhadas até a granularidade horária, com todas as dimensões possíveis.

Você pode usar ferramentas nativas da AWS, como o AWS Cost Explorer, AWS Budgetse aos Amazon Athena com o Amazon QuickSight para fornecer esse recurso. Você também pode usar ferramentas de terceiros. No entanto, você deve garantir que os custos dessas ferramentas forneçam valor à sua organização.

Etapas da implementação

- Crie um grupo de otimização de custos: Configure sua conta e crie um grupo que tenha acesso aos relatórios de custos e uso necessários. Esse grupo deve incluir representantes de todas as equipes que têm ou gerenciam um aplicativo. Isso garante que cada equipe tenha acesso às próprias informações de custo e uso.
- Configure o AWS Budgets: Configure o AWS Budgets em todas as contas para a sua workload. Defina um orçamento para o gasto total da conta e outro para a carga de trabalho usando tags.
- Configure o AWS Cost Explorer: Configure o AWS Cost Explorer para sua workload e contas. Crie um painel para a carga de trabalho que monitora o gasto geral e as principais métricas de uso da carga de trabalho.

- Configure ferramentas avançadas: Como opção, você pode criar ferramentas personalizadas para sua organização que fornecem detalhes e granularidade adicionais. Você pode implementar o recurso de análise avançada usando o Amazon Athenae painéis usando o Amazon QuickSight.

Recursos

Documentos relacionados:

- Marcação de recursos da AWS
- Análise de custos com o AWS Budgets
- Análise de custos com o Cost Explorer
- Gerenciamento do Relatório de Custos e Uso da AWS

Exemplos relacionados:

- Laboratórios do Well-Architected: Configuração da conta da AWS
- Laboratórios do Well-Architected: Visualização do faturamento
- Laboratórios do Well-Architected: Governança de custo e uso
- Laboratórios do Well-Architected: Análise de custo e uso
- Laboratórios do Well-Architected: Visualização de custo e uso

COST03-BP05 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.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

Implemente a marcação na AWS para adicionar informações da organização aos seus recursos, que serão adicionadas às suas informações de custo e uso. Uma tag é um par de chave-valor - a chave é definida e deve ser exclusiva em toda a organização, e o valor é exclusivo para um grupo de recursos. Um exemplo de um par de chave-valor é a chave Environment, com um valor de Production. Todos os recursos no ambiente de produção terão esse par de chave-valor. A marcação permite categorizar e rastrear seus custos com informações relevantes e significativas

da organização. Você pode aplicar tags que representem categorias da organização (como centros de custo, nomes de aplicação, projetos ou proprietários) e identificar workloads e características de workloads (como teste ou produção) para atribuir seus custos e uso em toda a organização.

Quando você aplica tags a seus recursos da AWS (como instâncias do Amazon Elastic Compute Cloud ou buckets do Amazon Simple Storage Service) e as ativa, a AWS adiciona essas informações aos Relatórios de Custo e Uso. Você pode gerar relatórios e realizar análises em recursos marcados e não marcados para permitir maior conformidade com políticas internas de gerenciamento de custos e garantir a atribuição precisa.

Criar e implementar um padrão de marcação da AWS em todas as contas da organização permite que você gerencie e administre seus ambientes da AWS de maneira consistente e uniforme. Use Políticas de tags no AWS Organizations para definir regras de como as tags podem ser usadas em recursos da AWS em suas contas no AWS Organizations. As políticas de tag permitem que você adote facilmente uma abordagem padronizada para marcar recursos da AWS.

AWS Tag Editor permite adicionar, excluir e gerenciar tags de vários recursos.

Categorias de Custo da AWS permite atribuir significado organizacional aos seus custos, sem exigir tags nos recursos. Você pode mapear suas informações de custo e uso para estruturas internas exclusivas da organização. Você define regras de categoria para mapear e categorizar custos usando dimensões de faturamento, como contas e tags. Isso fornece outro nível de capacidade de gerenciamento, além da marcação. Você também pode mapear contas e tags específicas para vários projetos.

Etapas da implementação

- Definir um esquema de marcação: Reúna todas as partes interessadas de toda a sua empresa para definir um esquema. Isso geralmente inclui pessoas dos departamentos técnico, financeiro e de gerenciamento. Defina uma lista de tags que todos os recursos devem ter, bem como outra lista com as tags que os recursos podem ter. Verifique se os nomes e valores das tags são consistentes em toda a organização.
- Marcar recursos: Usando suas categorias de atribuição de custo definidas, coloque tags em todos os recursos em suas cargas de trabalho de acordo com as categorias. Use ferramentas como CLI, Tag Editor ou Systems Manager para aumentar a eficiência.
- Implementar Categorias de Custos: Você pode criar categorias de custo sem implementar a marcação. As categorias de custos usam as dimensões de custo e uso existentes. Crie regras de categoria com base no esquema e as implemente no Categorias de Custos.

- Automatizar a marcação: Para garantir que você mantenha altos níveis de marcação em todos os recursos, automatize a marcação para que os recursos sejam marcados automaticamente quando forem criados. Use os recursos dentro do serviço ou use serviços como o AWS CloudFormation para garantir que os recursos sejam marcados quando criados. Você também pode criar um microsserviço personalizado que verifica a carga de trabalho periodicamente e remove todos os recursos que não estão marcados, o que é ideal para ambientes de teste e desenvolvimento.
- Monitorar e gerar relatórios sobre marcação: Para garantir que você mantenha altos níveis de marcação em toda a organização, relate e monitore as tags em todas as workloads. Você pode usar o AWS Cost Explorer para visualizar o custo de recursos marcados e não marcados ou usar serviços como o Tag Editor. Analise regularmente o número de recursos não marcados com tags e tome medidas para adicionar tags até atingir o nível desejado de marcação.

Recursos

Documentos relacionados:

- Tag de recurso do AWS CloudFormation
- Categorias de Custo da AWS
- Marcação de recursos da AWS
- O Amazon EC2 e o Amazon EBS incluem suporte para a marcação de recursos na criação
- Análise de custos com o AWS Budgets
- Análise de custos com o Cost Explorer
- Gerenciamento do Relatório de Custos e Uso da AWS

COST03-BP06 Alocar custos com base nas métricas de workload

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.

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

Orientação de implementação

a otimização de custos está fornecendo resultados de negócios com o menor preço, que só pode ser alcançado ao alocar custos de carga de trabalho por métricas de carga de trabalho (medidas

pela eficiência da carga de trabalho). Monitore as métricas de carga de trabalho definidas por meio de arquivos de log ou outro monitoramento de aplicativos. Combine esses dados com os custos da carga de trabalho, que podem ser obtidos examinando os custos com um valor de tag específico ou ID de conta. É recomendável executar essa análise no nível por hora. Sua eficiência normalmente mudará se você tiver alguns componentes de custo estático (por exemplo, um banco de dados de back-end em execução 24 horas por dia, 7 dias por semana) com uma taxa de solicitações variável (por exemplo, picos de uso entre 9h e 17h, com poucas solicitações à noite). Entender a relação entre os custos estáticos e variáveis ajudará você a concentrar suas atividades de otimização.

Etapas da implementação

- Alocar custos para métricas de workload: Usando as métricas definidas e a marcação configurada, crie uma métrica que combine a saída e o custo da carga de trabalho. Use os serviços de estudo analítico, como o Amazon Athena e o Amazon QuickSight, para criar um painel de eficiência para a workload geral e todos os componentes.

Recursos

Documentos relacionados:

- Marcação de recursos da AWS
- Análise de custos com o AWS Budgets
- Análise de custos com o Cost Explorer
- Gerenciamento do Relatório de Custos e Uso da AWS

COST 4 Como você desativa 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.

Práticas recomendadas

- COST04-BP01 Acompanhar os recursos ao longo da vida útil
- COST04-BP02 Implemente um processo de desativação
- COST04-BP03 Desativar recursos
- COST04-BP04 Desative recursos automaticamente

COST04-BP01 Acompanhar os recursos ao longo da vida útil

Defina e implemente um método para acompanhar recursos e suas associações com sistemas ao longo da vida útil. Você pode usar a marcação para identificar a carga de trabalho ou a função do recurso.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

desative recursos de carga de trabalho que não são mais necessários. Um exemplo comum são os recursos usados para testes, após a conclusão do teste, os recursos podem ser removidos. Rastrear recursos com tags (e executar relatórios sobre essas tags) ajudará você a identificar ativos para desativação. Usar tags é uma maneira eficaz de rastrear recursos, rotulando o recurso com sua função ou uma data conhecida em que ele pode ser desativado. Os relatórios podem ser executados nessas tags. Os valores de exemplo para marcação de recursos são `testes de featureX` para identificar a finalidade do recurso em termos de ciclo de vida da workload.

Etapas da implementação

- Implemente um esquema de marcação: Implemente um esquema de marcação que identifique a workload à qual o recurso pertence, garantindo que todos os recursos dentro da workload sejam marcados da maneira apropriada.
- Implemente o monitoramento da saída ou do throughput da workload: Implemente o alarme ou monitoramento do throughput da carga de trabalho, acionando solicitações de entrada ou conclusões de saída. Configure-o para fornecer notificações quando saídas ou solicitações de carga de trabalho caírem para zero, indicando que os recursos de carga de trabalho não são mais usados. Incorpore um fator de tempo se a carga de trabalho cair periodicamente para zero em condições normais.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor
- Marcação de recursos da AWS
- Publicar métricas personalizadas

COST04-BP02 Implemente um processo de desativação

Implemente um processo para identificar e desativar recursos órfãos.

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

Orientação de implementação

implemente um processo padronizado em toda a organização para identificar e remover recursos não utilizados. O processo deve definir a frequência das pesquisas e os processos para remover o recurso para garantir que todos os requisitos da organização sejam atendidos.

Etapas da implementação

- Crie e implemente um processo de desativação: Trabalhando com os proprietários e desenvolvedores de cargas de trabalho, crie um processo de desativação para a carga de trabalho e os recursos dela. O processo deve abranger o método para verificar se a carga de trabalho está em uso e também se cada um dos recursos da carga de trabalho está em uso. O processo também deve abranger as etapas necessárias para desativar o recurso, removendo-os do serviço e garantindo a conformidade com os requisitos normativos. Todos os recursos associados, como licenças ou armazenamento anexado, também são cobertos. Por fim, o processo deve fornecer uma notificação aos proprietários da workload de que o processo de desativação foi executado.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor

COST04-BP03 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.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

a frequência e o esforço para pesquisar recursos não utilizados devem refletir as possíveis economias, portanto, uma conta com um custo pequeno deve ser analisada com menos frequência

do que uma conta com custos maiores. Pesquisas e eventos de desativação podem ser acionados por alterações de estado na carga de trabalho, como um produto que termina a vida útil ou é substituído. Pesquisas e eventos de desativação também podem ser acionados por eventos externos, como alterações nas condições de mercado ou encerramento do produto.

Etapas da implementação

- Desativar recursos: Usando o processo de desativação, desative cada um dos recursos que foram identificados como órfãos.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor

COST04-BP04 Desative 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.

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

Orientação de implementação

use a automação para reduzir ou remover os custos associados do processo de desativação. Projetar sua carga de trabalho para executar a desativação automatizada reduzirá os custos gerais da carga de trabalho durante sua vida útil. Você pode usar o AWS Auto Scaling para executar o processo de desativação. Você também pode implementar código personalizado usando a API ou o SDK para desativar recursos de carga de trabalho automaticamente.

Etapas da implementação

- Implemente o AWS Auto Scaling: Configure os recursos compatíveis com o AWS Auto Scaling.
- Configure o CloudWatch para encerrar instâncias: As instâncias podem ser configuradas para serem encerradas usando alarmes do CloudWatch. Usando as métricas do processo de desativação, implemente um alarme com uma ação do Amazon Elastic Compute Cloud (Amazon EC2). Verifique a operação em um ambiente que não seja de produção antes de implantar.

- Implemente código dentro da workload: Você pode usar o SDK ou o AWS CLI da AWS para desativar recursos de workload. Implemente código dentro da aplicação que se integra à AWS e encerre ou remova recursos não mais usados.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor
- Crie alarmes para interromper, encerrar, reinicializar ou recuperar uma instância
- Conceitos básicos do Amazon EC2 Auto Scaling

Recursos econômicos

Perguntas

- COST 5 Como você avalia o custo ao selecionar serviços?
- COST 6 Como você atinge as metas de custo ao selecionar tamanho, número e tipo de recurso?
- COST 7 Como você usa modelos de definição de preço para reduzir custos?
- COST 8 Como você planeja as cobranças de transferência de dados?

COST 5 Como você avalia o custo ao selecionar serviços?

O Amazon EC2, o Amazon EBS e o Amazon S3 são serviços fundamentais da AWS. Serviços gerenciados como o Amazon RDS e o Amazon DynamoDB são serviços da AWS de nível superior ou em nível de aplicação. 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.

Práticas recomendadas

- COST05-BP01 Identificar requisitos da organização para custos
- COST05-BP02 Analisar todos os componentes desta workload
- COST05-BP03 Executar uma análise completa de cada componente

- COST05-BP04 Selecionar software com licenciamento econômico
- COST05-BP05 Selecionar os componentes dessa workload para otimizar o custo alinhado com as prioridades da organização
- COST05-BP06 Realize análises de custos para diferentes usos ao longo do tempo

COST05-BP01 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.

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

Orientação de implementação

ao selecionar serviços para a sua carga de trabalho, é fundamental compreender as prioridades da sua organização. Verifique se você tem um equilíbrio entre custo e outros pilares do Well-Architected, como performance e confiabilidade. Uma carga de trabalho totalmente otimizada para custo é a solução mais alinhada aos requisitos da sua organização, não necessariamente o menor custo. Reúna-se com todas as equipes da sua organização para coletar informações, como produtos, negócios, técnicas e finanças.

Etapas da implementação

- Identificar requisitos da organização para custos: Reúna-se com membros da equipe da sua organização, incluindo aqueles em gerenciamento de produtos, proprietários de aplicações, equipes de desenvolvimento e operações, gerenciamento e finanças. Priorize os pilares do Well-Architected para essa workload e os componentes correspondentes; o resultado é uma lista dos pilares em ordem. Você também pode adicionar uma ponderação a cada, o que pode ajudar a indicar quanto foco adicional um pilar tem ou a semelhança do foco entre dois pilares.

Recursos

Documentos relacionados:

- Calculadora de custo total de propriedade (TCO) da AWS
- Classes de armazenamento do Amazon S3
- Produtos da nuvem

COST05-BP02 Analisar todos os componentes desta workload

Garanta que todos os componentes da workload sejam analisados, independentemente do tamanho atual ou dos custos atuais. O esforço da análise deve refletir o benefício potencial, como custos atuais e projetados.

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

Orientação de implementação

faça uma análise completa de todos os componentes na sua carga de trabalho. Garantir o equilíbrio entre o custo da análise e as possíveis economias na carga de trabalho durante seu ciclo de vida. Você deve encontrar o impacto atual e o possível impacto futuro do componente. Por exemplo, se o custo do recurso proposto for de USD 10 por mês, e sob as cargas previstas não excederem USD 15 por mês, gastar um dia de esforço para reduzir custos em 50% (USD 5 por mês) poderá exceder o possível benefício durante a vida útil do sistema. Usar uma estimativa baseada em dados mais rápida e eficiente criará o melhor resultado geral para esse componente.

As workloads podem mudar ao longo do tempo. O conjunto certo de serviços pode não ser ideal se a arquitetura da workload ou o uso mudar. A análise para seleção de serviços deve incorporar estados de carga de trabalho e níveis de uso atuais e futuros. A implementação de um serviço para o estado ou uso futuro da carga de trabalho pode reduzir os custos gerais ao reduzir ou remover o esforço necessário para fazer alterações futuras.

AWS Cost Explorer e a seção AWS Cost and Usage Report (CUR) podem analisar o custo de uma prova de conceito (PoC) ou de um ambiente em execução. Você também pode usar o AWS Pricing Calculator para estimar os custos da carga de trabalho.

Etapas da implementação

- Listar os componentes da workload: Crie a lista de todos os componentes da workload. Ela é usada como verificação para conferir se cada componente foi analisado. O esforço gasto deve refletir a criticidade da workload conforme definido pelas prioridades da sua organização. Agrupar recursos funcionalmente melhora a eficiência, por exemplo, o armazenamento do banco de dados de produção, se houver vários bancos de dados.
- Priorizar a lista de componentes: Pegue a lista de componentes e a priorize em ordem de esforço. Isso geralmente ocorre na ordem do custo do componente, do mais caro para o mais barato, ou da criticidade, conforme definido pelas prioridades da organização.
- Executar a análise: Para cada componente na lista, analise as opções e os serviços disponíveis e escolha a opção mais alinhada com suas prioridades organizacionais.

Recursos

Documentos relacionados:

- AWS Pricing Calculator
- AWS Cost Explorer
- Classes de armazenamento do Amazon S3
- Produtos da nuvem

COST05-BP03 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.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

Considere a economia de tempo que permitirá que sua equipe se concentre na aposentadoria de recursos de endividamento técnico, inovação e agregação de valor. Por exemplo, talvez você precise mover sem alterações (lift and shift) rapidamente seu ambiente on-premises para a nuvem e otimizar mais tarde. Vale a pena explorar as economias que você poderia obter usando serviços gerenciados que removem ou reduzem os custos de licença. serviços gerenciados eliminam a sobrecarga operacional e administrativa da manutenção de um serviço, o que permite que você se concentre na inovação. Além disso, como serviços gerenciados operam em escala da nuvem, eles podem oferecer menor custo por transação ou serviço.

Geralmente, os serviços gerenciados têm atributos que você pode definir para garantir capacidade suficiente. Você deve definir e monitorar esses atributos para que sua capacidade em excesso seja mínima e a performance seja maximizada. Você pode modificar os atributos do AWS Managed Services usando o AWS Management Console ou as APIs e os SDKs da AWS para alinhar as necessidades de recursos com a demanda em constante mudança. Por exemplo, você pode aumentar ou diminuir o número de nós em um cluster do Amazon EMR (ou em um cluster do Amazon Redshift) para aumentar ou reduzir a escala horizontalmente.

Você também pode unir várias instâncias em um recurso da AWS para habilitar usos de maior densidade. Por exemplo, você pode provisionar vários bancos de dados pequenos em uma única

instância de banco de dados do Amazon Relational Database Service (Amazon RDS). Conforme o uso aumenta, você pode migrar um dos bancos de dados para uma instância de banco de dados Amazon RDS dedicada usando um processo de snapshot e restauração.

Ao provisionar cargas de trabalho em serviços gerenciados, você deve compreender os requisitos de ajuste da capacidade do serviço. Esses requisitos geralmente são tempo, esforço e qualquer impacto na operação normal da carga de trabalho. O recurso provisionado deve permitir tempo para que as alterações ocorram, provisionar a sobrecarga necessária para permitir isso. O trabalho contínuo necessário para modificar os serviços pode ser reduzido a praticamente zero usando APIs e SDKs integrados a ferramentas de sistema e monitoramento como o Amazon CloudWatch.

Amazon RDS, Amazon Redshifte aos Amazon ElastiCache fornecem um serviço de banco de dados gerenciado. Amazon Athena, Amazon EMRe aos Amazon OpenSearch Service fornecem um serviço de análise gerenciado.

AMS é um serviço que opera a infraestrutura da AWS em nome de clientes e parceiros empresariais. Ele fornece um ambiente seguro e compatível no qual você pode implantar suas workloads. O AMS usa modelos operacionais de nuvem empresarial com automação para permitir que você atenda aos requisitos da sua organização, migre para a nuvem mais rapidamente e reduza seus custos de gerenciamento constantes.

Etapas da implementação

- Executar uma análise completa: Usando a lista de componentes, trabalhe com cada componente da maior prioridade para a menor. Para componentes de prioridade maior e mais caros, execute análises adicionais e avalie todas as opções disponíveis e o impacto a longo prazo. Para componentes de prioridade menor, avalie se alterações no uso alterariam a prioridade do componente e, em seguida, execute uma análise de esforço apropriado.

Recursos

Documentos relacionados:

- Calculadora de custo total de propriedade (TCO) da AWS
- Classes de armazenamento do Amazon S3
- Produtos da nuvem

COST05-BP04 Selecionar software com licenciamento econômico

Os softwares de código aberto eliminarão os custos de licenciamento de software, o que pode contribuir com custos significativos para as workloads. 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.

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

Orientação de implementação

O custo das licenças de software pode ser eliminado com o uso de software de código aberto. Isso pode ter impacto significativo nos custos da carga de trabalho à medida que o tamanho da carga de trabalho é dimensionado. Meça os benefícios do software licenciado em relação ao custo total para garantir que você tenha a carga de trabalho mais otimizada. Modele todas as alterações no licenciamento e como elas afetariam seus custos de carga de trabalho. Se um fornecedor alterar o custo da sua licença de banco de dados, investigue como isso afeta a eficiência geral da sua carga de trabalho. Considere anúncios históricos de definição de preço de seus fornecedores para tendências de alterações de licenciamento em seus produtos. Os custos de licenciamento também podem ser dimensionados independentemente do throughput ou do uso, como licenças que escalam por hardware (licenças vinculadas à CPU). Essas licenças devem ser evitadas porque os custos podem aumentar rapidamente sem resultados correspondentes.

Etapas da implementação

- Analisar opções de licença: Analise os termos de licenciamento do software disponível. Procure versões de código aberto que tenham a funcionalidade necessária e veja se os benefícios do software licenciado superam o custo. Termos favoráveis alinham o custo do software aos benefícios que ele oferece.
- Analisar o provedor de software: Analise todas as alterações históricas de definição de preço ou licenciamento do fornecedor. Procure alterações que não estejam alinhadas aos resultados, como termos punitivos para execução em hardware ou plataformas de fornecedores específicos. Além disso, procure como eles executam auditorias e penalidades que poderiam ser impostas.

Recursos

Documentos relacionados:

- Calculadora de custo total de propriedade (TCO) da AWS
- Classes de armazenamento do Amazon S3
- Produtos da nuvem

COST05-BP05 Selecionar os componentes dessa workload para otimizar o custo alinhado com as prioridades da organização

Considere o custo ao selecionar todos os componentes. Isso inclui o uso de serviços gerenciados e em nível de aplicação, como o Amazon Relational Database Service (Amazon RDS), Amazon DynamoDB, Amazon Simple Notification Service (Amazon SNS) e Amazon Simple Email Service (Amazon SES) para reduzir o custo geral da organização. Use contêineres e recursos de tecnologia sem servidor para computação, como o AWS Lambda, o Amazon Simple Storage Service (Amazon S3) para sites estáticos, e o Amazon Elastic Container Service (Amazon ECS). Minimize os custos de licença usando software de código aberto ou software sem taxas de licença: por exemplo, Amazon Linux para workloads de computação ou migração de bancos de dados para o Amazon Aurora.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

você pode usar serviços sem servidor ou em nível de aplicativo, como o AWS Lambda, o Amazon Simple Queue Service (Amazon SQS), Amazon SNSe aos Amazon SES. Esses serviços eliminam a necessidade de gerenciar um recurso e fornecem a função de execução de código, serviços de enfileiramento e entrega de mensagens. O outro benefício é que eles escalam a performance e o custo de acordo com o uso, permitindo a alocação e a atribuição eficientes de custos.

Para obter mais informações sobre o recurso Serverless, consulte o whitepaper Well-Architected Serverless Application Lens..

Etapas da implementação

- Selecionar cada serviço para otimizar o custo: Usando sua análise e lista priorizada, selecione cada opção que fornece a melhor correspondência com suas prioridades organizacionais.

Recursos

Documentos relacionados:

- Calculadora de custo total de propriedade (TCO) da AWS
- Classes de armazenamento do Amazon S3
- Produtos da nuvem

COST05-BP06 Realize 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, a workload continua oferecendo um bom custo-benefício ao longo da vida útil.

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

Orientação de implementação

À medida que a AWS lança novos serviços e recursos, os serviços ideais para sua workload podem mudar. O esforço necessário deve refletir possíveis benefícios. A frequência da análise da carga de trabalho depende dos requisitos da sua organização. Se for uma carga de trabalho com custo significativo, implementar novos serviços mais cedo maximizará a economia de custos, portanto, uma revisão mais frequente poderá ser vantajosa. Outro trigger para revisão é a alteração nos padrões de uso. Alterações significativas no uso podem indicar que serviços alternativos seriam mais ideais. Por exemplo, para taxas de transferência de dados mais altas, um serviço de conexão direta pode ser mais barato do que uma VPN e fornecer a conectividade necessária. Preveja o possível impacto das alterações de serviço para que você possa monitorar esses acionadores de nível de uso e implementar os serviços mais econômicos mais cedo.

Etapas da implementação

- Defina padrões de uso previstos: Trabalhando com sua organização, como proprietários de produtos e marketing, documente quais serão os padrões de uso previstos e esperados para a workload.
- Execute análise de custos no uso previsto: Usando os padrões de uso definidos, execute a análise em cada um desses pontos. O esforço de análise deve refletir o resultado provável. Por exemplo, se a alteração no uso for grande, uma análise completa deverá ser realizada para verificar quaisquer custos e alterações.

Recursos

Documentos relacionados:

- Calculadora de custo total de propriedade (TCO) da AWS
- Classes de armazenamento do Amazon S3
- Produtos da nuvem

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.

Práticas recomendadas

- COST06-BP01 Executar modelagem de custos
- COST06-BP02 Selecione o tipo, tamanho e número do recurso com base nos dados
- COST06-BP03 Selecionar o tipo, tamanho e número do recurso automaticamente com base nas métricas

COST06-BP01 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.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Alto

Orientações para a implementação

Execute a modelagem de custos para sua carga de trabalho e cada um de seus componentes para entender o equilíbrio entre recursos e encontrar o tamanho correto para cada recurso na carga de trabalho, dado um nível específico de performance. 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 ou à economia prevista. Para conhecer as práticas recomendadas, consulte a seção Análise do Whitepaper sobre pilar de eficiência de performance.

AWS Compute Optimizer pode ajudar na modelagem de custos para a execução de workloads. Ele fornece recomendações de dimensionamento correto para recursos de computação com base no uso histórico. Essa é a fonte de dados ideal para recursos de computação, pois é um serviço gratuito e utiliza Machine Learning para fazer várias recomendações, dependendo dos níveis de risco. Você também pode usar o Amazon CloudWatch e Amazon CloudWatch Logs com logs personalizados como fontes de dados para operações de dimensionamento correto para outros serviços e componentes de carga de trabalho.

Veja a seguir as recomendações para dados e métricas de modelagem de custo:

- O monitoramento deve refletir com precisão a experiência do usuário final. Selecione a granularidade correta para o período e escolha com cuidado o máximo ou o 99º percentil, em vez da média.
- Selecione a granularidade correta para o período de análise necessário para cobrir todos os ciclos de carga de trabalho. Por exemplo, se uma análise de duas semanas for realizada, talvez você esteja deixando passar um ciclo de alta utilização, o que pode levar a subprovisionamento.

Etapas da implementação

- Executar modelagem de custos: Implante a workload ou uma prova de conceito em uma conta separada com os tipos e tamanhos de recursos específicos a serem testados. Execute a workload com os dados de teste e registre os resultados de saída, juntamente com os dados de custo da hora em que o teste foi executado. Em seguida, reimplante a workload ou altere os tipos e tamanhos de recursos e execute novamente o teste.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- Recursos do Amazon CloudWatch
- Otimização de custos: dimensionamento correto do Amazon EC2
- AWS Compute Optimizer

COST06-BP02 Selecione o tipo, tamanho e número do recurso com base nos dados

Selecione o tamanho ou tipo de recurso com base nos dados sobre a workload e nas características do recurso. Por exemplo, computação, memória, throughput ou gravação intensiva. Essa seleção geralmente é feita usando uma versão anterior (on-premises) da workload, usando a documentação ou outras fontes de informações sobre a workload.

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

Orientação de implementação

Selecione o tamanho ou tipo de recurso com base na workload e nas características do recurso, por exemplo, computação, memória, throughput ou gravação intensiva. Essa seleção geralmente é feita usando a modelagem de custo, uma versão anterior da workload (como uma versão on-premises), usando a documentação ou outras fontes de informações sobre a workload (whitepapers, soluções publicadas).

Etapas da implementação

- Selecione recursos com base em dados: Usando seus dados de modelagem de custo, selecione o nível de uso esperado da workload e, em seguida, o tipo e o tamanho do recurso especificado.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- Recursos do Amazon CloudWatch
- Otimização de custos: dimensionamento correto do EC2

COST06-BP03 Selecionar o tipo, tamanho e número 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. Provisione adequadamente o throughput, o dimensionamento e o armazenamento para serviços como o Amazon Elastic Compute Cloud (Amazon EC2), o Amazon DynamoDB, o Amazon Elastic Block Store (Amazon EBS) (PIOPS), o Amazon Relational Database Service (Amazon RDS), o Amazon EMR e redes. Isso pode ser feito com um ciclo de comentários, como escalabilidade automática ou por código personalizado na carga de trabalho.

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

Orientação de implementação

crie um loop de comentários dentro da carga de trabalho que usa métricas ativas da carga de trabalho em execução para fazer alterações nessa carga de trabalho. É possível usar um serviço gerenciado, como o AWS Auto Scaling, que você configura para realizar as operações de dimensionamento certas. A AWS também oferece APIs, SDKse funcionalidades que permitem que os recursos sejam modificados com o mínimo de esforço. É possível programar uma workload para interromper e iniciar uma instância do Amazon Elastic Compute Cloud(Amazon EC2) para permitir uma alteração de tamanho ou tipo de instância. Isso fornece os benefícios do dimensionamento correto e, ao mesmo tempo, remove quase todo o custo operacional necessário para fazer a alteração.

Alguns serviços da AWS têm seleção automática de tipo ou tamanho, como o Amazon Simple Storage Service(Amazon S3) Intelligent-Tiering. O Amazon S3 Intelligent-Tiering move automaticamente seus dados entre dois níveis de acesso: acesso frequente e acesso infrequente, com base em seus padrões de uso.

Etapas da implementação

- Configurar métricas de workload: Capture as principais métricas para a carga de trabalho. Essas métricas fornecem uma indicação da experiência do cliente, como a saída da carga de trabalho, e se alinham às diferenças entre tipos e tamanhos de recursos, como uso de CPU e memória.
- Visualizar recomendações de dimensionamento correto: Use as recomendações de dimensionamento correto no AWS Compute Optimizer para fazer ajustes na workload.
- Selecionar o tipo e o tamanho do recurso automaticamente com base nas métricas: Usando as métricas de carga de trabalho, selecione de modo manual ou automático seus recursos de carga de trabalho. A configuração do AWS Auto Scaling ou a implementação de código dentro da aplicação pode reduzir o esforço necessário se alterações frequentes forem necessárias e, possivelmente, implementar alterações antes de um processo manual.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Compute Optimizer

- Recursos do Amazon CloudWatch
- Como configurar o CloudWatch
- Publicar métricas personalizadas do CloudWatch
- Otimização de custos: dimensionamento correto do Amazon EC2
- Conceitos básicos do Amazon EC2 Auto Scaling
- Amazon S3 Intelligent-Tiering
- Iniciar uma instância do EC2 usando o SDK

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

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

Práticas recomendadas

- COST07-BP01 Executar análise de modelo de definição de preço
- COST07-BP02 Implementar regiões com base nos custos
- COST07-BP03 Selecionar contratos de terceiros com termos econômicos
- COST07-BP04 Implemente modelos de definição de preço para todos os componentes dessa workload
- COST07-BP05 Execute a análise do modelo de definição de preço no nível da conta mestre

COST07-BP01 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 instâncias sob demanda ou spot). Execute uma análise da carga de trabalho usando o recurso Recomendações no AWS Cost Explorer.

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

Orientação de implementação

A AWS tem vários modelos de definição de preço que permitem que você pague pelos seus recursos da maneira mais econômica que atenda às necessidades da sua organização.

Etapas da implementação

- Executar uma análise de desconto de compromisso: Usando o Cost Explorer em sua conta, analise as recomendações de instâncias reservadas e Savings Plans. Para garantir que você implemente as recomendações corretas com os descontos e riscos necessários, siga os Laboratórios do Well-Architected.
- Analisar a elasticidade da workload: Usando a granularidade por hora no Cost Explorer ou um painel personalizado. Analise a elasticidade da workload. Procure alterações regulares no número de instâncias em execução. As instâncias de curta duração são candidatas a instâncias spot ou frota spot.
  - Laboratório do Well-Architected: Cost Explorer
  - Laboratório do Well-Architected: Visualização de custos

Recursos

Documentos relacionados:

- Acesso a recomendações de instância reservada
- Opções de compra de instância

Vídeos relacionados:

- Economize até 90% e execute cargas de trabalho de produção no local

Exemplos relacionados:

- Laboratório do Well-Architected: Cost Explorer
- Laboratório do Well-Architected: Visualização de custos
- Laboratório do Well-Architected: Modelos de definição de preço

COST07-BP02 Implementar regiões com base nos custos

A definição de preço dos recursos pode ser diferente em cada região. A consideração do custo da região ajuda a garantir que você pague o menor preço geral por essa workload.

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

Orientação de implementação

Ao projetar suas soluções, uma melhor prática é buscar colocar os recursos de computação mais perto dos usuários para proporcionar menor latência e forte soberania de dados. Para públicos globais, você deve usar vários locais para atender a essas necessidades. Você deve selecionar a localização geográfica que minimiza seus custos.

A infraestrutura da Nuvem AWS é criada com base em Regiões e zonas de disponibilidade. Região é um local físico do mundo onde há várias zonas de disponibilidade. As zonas de disponibilidade consistem em um ou mais datacenters discretos que estão alojados em instalações separadas, cada uma com energia, rede e conectividade redundantes.

Cada região da AWS opera dentro das condições do mercado local, e a definição de preço dos recursos é diferente em cada região. Escolha uma região específica para operar um componente de sua solução completa para que você possa operar ao menor preço possível globalmente. Você pode usar o AWS Pricing Calculator para estimar os custos da workload em várias regiões.

Etapas da implementação

- Analisar a definição de preço da região: Analise os custos da workload na região atual. Começando com os custos maiores por serviço e tipo de uso, calcule os custos em outras regiões que estão disponíveis. Se a economia prevista ultrapassar o custo de mover o componente ou a workload, migre para a nova região.

Recursos

Documentos relacionados:

- Acesso a recomendações de instância reservada
- Definição de preço do Amazon EC2
- Opções de compra de instância
- Tabela de regiões

Vídeos relacionados:

- Economize até 90% e execute cargas de trabalho de produção no local

COST07-BP03 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.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

quando você utiliza soluções ou serviços de terceiros na nuvem, é importante que as estruturas de definição de preço estejam alinhadas aos resultados da otimização de custos. A definição de preço deve ser dimensionada de acordo com os resultados e o valor que fornece. Um exemplo disso é um software que leva uma porcentagem das economias que ele fornece, quanto mais você economiza (resultado), mais ele cobra. Contratos que escalam com sua fatura normalmente não estão alinhados com a otimização de custos, a menos que forneçam resultados para cada parte da sua fatura específica. Por exemplo, uma solução que fornece recomendações para o Amazon Elastic Compute Cloud(Amazon EC2) e cobra uma porcentagem de toda a sua fatura aumentará se você usar outros serviços para os quais ela não oferece nenhum benefício. Outro exemplo é um serviço gerenciado que é cobrado a uma porcentagem do custo dos recursos que são gerenciados. Um tamanho de instância maior pode não exigir necessariamente mais esforço de gerenciamento, mas será cobrado mais. Certifique-se de que essas disposições de definição de preço de serviços incluam um programa de otimização de custos ou recursos em seu serviço para promover a eficiência.

Etapas da implementação

- Analisar contratos e termos de terceiros: Analise a definição de preço em contratos de terceiros. Execute modelagem para diferentes níveis de uso e inclua novos custos, como uso de novos serviços, ou aumentos nos serviços atuais devido ao crescimento da workload. Decida se os custos adicionais fornecem os benefícios necessários para a sua empresa.

Recursos

Documentos relacionados:

- Acesso a recomendações de instância reservada
- Opções de compra de instância

Vídeos relacionados:

- Economize até 90% e execute cargas de trabalho de produção no local

COST07-BP04 Implemente modelos de definição de preço para todos os componentes dessa workload

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. As instâncias sob demanda são usadas somente para workloads 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.

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

Orientação de implementação

considere os requisitos dos componentes da carga de trabalho e entenda os possíveis modelos de definição de preço. Defina o requisito de disponibilidade do componente. Determine se há vários recursos independentes que executam a função na carga de trabalho e quais são os requisitos da carga de trabalho ao longo do tempo. Compare o custo dos recursos usando o modelo de definição de preço sob demanda padrão e outros modelos aplicáveis. Leve em consideração possíveis alterações nos recursos ou componentes da carga de trabalho.

Etapas da implementação

- Implemente modelos de definição de preço: Usando seus resultados de análise, compre Savings Plans (SPs), instâncias reservadas (RIs) ou implemente instâncias spot. Se for sua primeira compra de RI, escolha as 5 ou 10 recomendações principais na lista e, em seguida, monitore e analise os resultados dos próximos dois meses, no máximo. Compre pequenas quantidades de descontos de compromisso em ciclos regulares, por exemplo, a cada duas semanas ou mensalmente. Implemente instâncias spot para workloads que podem ser interrompidas ou são sem estado.
- Ciclo de análise da workload: Implemente um ciclo de análise da carga de trabalho que analise especificamente a cobertura do modelo de definição de preço. Assim que a workload tiver a cobertura necessária, compre descontos de compromisso adicionais a cada 2 a 4 semanas ou conforme o uso da organização mudar.

Recursos

Documentos relacionados:

- Acesso a recomendações de instância reservada
- Frota EC2
- Como comprar instâncias reservadas
- Opções de compra de instância
- Instâncias spot

Vídeos relacionados:

- Economize até 90% e execute cargas de trabalho de produção no local

COST07-BP05 Execute 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 de gerenciamento e obter descontos de compromisso.

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

Orientação de implementação

a execução de uma modelagem de custo regular garante que as oportunidades de otimização em várias cargas de trabalho possam ser implementadas. Por exemplo, se várias workloads usarem instâncias sob demanda, em um nível agregado, o risco de alteração será menor, e a implementação de um desconto baseado em compromisso atingirá um custo geral mais baixo. É recomendável realizar análises em ciclos regulares de duas semanas a um mês. Isso permite que você faça pequenas compras de ajuste, para que a cobertura de seus modelos de definição de preço continue a evoluir com suas cargas de trabalho dinâmicas e seus componentes.

Use a ferramenta de recomendações do AWS Cost Explorer para encontrar oportunidades de descontos de compromisso.

Para encontrar oportunidades para cargas de trabalho spot, use uma visualização por hora do uso geral e procure períodos regulares de uso ou elasticidade variáveis.

Etapas da implementação

- Executar uma análise de desconto de compromisso: Usando o Cost Explorer em sua conta, analise as recomendações de instâncias reservadas e Savings Plans. Para garantir que você implemente as recomendações corretas com os descontos e riscos necessários, siga os laboratórios do Well-Architected.

Recursos

Documentos relacionados:

- Acesso a recomendações de instância reservada
- Opções de compra de instância

Vídeos relacionados:

- Economize até 90% e execute cargas de trabalho de produção no local

Exemplos relacionados:

- Laboratório do Well-Architected: modelos de definição de preço

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.

Práticas recomendadas

- COST08-BP01 Execute modelagem de transferência de dados
- COST08-BP02 Selecione componentes para otimizar o custo de transferência de dados
- COST08-BP03 Implementar serviços para reduzir custos de transferência de dados

COST08-BP01 Execute 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.

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

Orientação de implementação

entenda onde a transferência de dados ocorre na carga de trabalho, o custo da transferência e o benefício associado. Isso permite que você tome uma decisão embasada para modificar ou

aceitar a decisão arquitetônica. Por exemplo, você pode ter uma configuração de várias zonas de disponibilidade na qual você replica dados entre as zonas de disponibilidade. Você modela o custo da estrutura e decide que esse é um custo aceitável (semelhante ao pagamento por computação e armazenamento em ambas as zonas de disponibilidade) para alcançar a confiabilidade e a resiliência necessárias.

Modele os custos em diferentes níveis de uso. O uso da carga de trabalho pode mudar ao longo do tempo, e diferentes serviços podem ser mais econômicos em diferentes níveis.

Uso AWS Cost Explorer ou o AWS Cost and Usage Report (CUR) para compreender e modelar seus custos de transferência de dados. Configure uma prova de conceito (PoC) ou teste sua carga de trabalho e execute um teste com uma carga simulada realista. Você pode modelar seus custos em diferentes demandas de carga de trabalho.

Etapas da implementação

- Calcular custos de transferência de dados: Use a ferramenta de recomendações do Páginas de definição de preço da AWS e calcule os custos de transferência de dados para a workload. Calcule os custos de transferência de dados em diferentes níveis de uso, tanto para aumentos quanto para reduções no uso da workload. Quando houver várias opções para o custo da arquitetura da workload, calcule o custo de cada uma delas para comparação.
- Vincular custos aos resultados: Para cada custo de transferência de dados incorrido, especifique o resultado que ele atinge para a carga de trabalho. Se for transferência entre componentes, poderá ser para desacoplamento; se estiver entre Zonas de Disponibilidade, poderá ser para redundância.

Recursos

Documentos relacionados:

- AWS caching solutions (Soluções de armazenamento em cache da AWS)
- Definição de preço da AWS
- Definição de preço da Amazon EC2
- Definição de preço da Amazon VPC
- Acelere a entrega de conteúdo com o Amazon CloudFront

COST08-BP02 Selecione 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 rede de longa distância (WAN) e configurações de várias zonas de disponibilidade (AZ).

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

Orientação de implementação

Criar a arquitetura para transferência de dados garante que você minimize os custos de transferência de dados. Isso pode envolver usar redes de entrega de conteúdo para colocar os dados mais perto dos usuários ou usar links de rede dedicados de seu local para a AWS. Você também pode usar a otimização de WAN e a otimização de aplicativos para reduzir a quantidade de dados transferidos entre componentes.

Etapas da implementação

- Selecionar componentes para transferência de dados: Usando a modelagem de transferência de dados, concentre-se em onde estão os maiores custos de transferência de dados ou onde estariam se o uso da workload mudasse. Procure arquiteturas alternativas ou componentes adicionais que eliminem ou reduzam a necessidade da transferência de dados ou reduzam o custo dela.

Recursos

Documentos relacionados:

- AWS caching solutions (Soluções de armazenamento em cache da AWS)
- Acelere a entrega de conteúdo com o Amazon CloudFront

COST08-BP03 Implementar serviços para reduzir custos de transferência de dados

Implemente serviços para reduzir a transferência de dados. Por exemplo, ao usar uma rede de entrega de conteúdo (CDN) como o Amazon CloudFront para fornecer conteúdo aos usuários finais, armazenar camadas em cache com o uso do Amazon ElastiCache ou usar o AWS Direct Connect no lugar da VPN para conectividade com a AWS.

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

Orientação de implementação

Amazon CloudFront é uma rede de entrega de conteúdo global que entrega dados com baixa latência e altas velocidades de transferência. Ele armazena dados em cache em pontos de presença no mundo inteiro, o que reduz a carga sobre seus recursos. Ao usar o CloudFront, você pode reduzir o trabalho administrativo para entregar conteúdo a grandes números de usuários globalmente com latência mínima.

AWS Direct Connect permite estabelecer uma conexão de rede dedicada com a AWS. Isso pode reduzir os custos de rede, aumentar a largura de banda e fornecer uma experiência de rede mais consistente do que conexões baseadas em Internet.

AWS VPN permite estabelecer uma conexão segura e privada entre sua rede privada e a rede global da AWS. Ele é ideal para pequenos escritórios ou parceiros de negócios porque oferece conectividade rápida e fácil, além de ser um serviço totalmente gerenciado e elástico.

VPC Endpoints permitem conectividade entre os serviços da AWS em redes privadas e podem ser usados para reduzir os custos de transferência de dados pública e gateways NAT. Endpoints da VPC de Gateway não têm cobranças por hora e oferecem suporte ao Amazon Simple Storage Service(Amazon S3) e ao Amazon DynamoDB. VPC endpoints de interface são fornecidos pelo AWS PrivateLink e têm uma taxa horária e por GB de custo para uso.

Etapas da implementação

- Implementar serviços: Usando a modelagem de transferência de dados, veja onde estão os maiores custos e os fluxos de volume mais altos. Analise os serviços da AWS e avalie se há um serviço que reduz ou remove a transferência, especificamente a entrega de conteúdo e redes. Procure também serviços de armazenamento em cache em que haja acesso repetido aos dados ou grandes quantidades de dados.

Recursos

Documentos relacionados:

- AWS Direct Connect
- Explore nossos produtos da AWS
- AWS caching solutions (Soluções de armazenamento em cache da AWS)
- Amazon CloudFront

- Acelere a entrega de conteúdo com com o Amazon CloudFront

Gerenciar recursos de demanda e fornecimento

Pergunta

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

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 em ambas as direções tem um impacto adverso sobre a organização, tanto nos custos operacionais (redução na performance em decorrência de utilização excessiva) quanto em despesas desnecessárias na AWS (devido ao excesso de provisionamento).

Práticas recomendadas

- COST09-BP01 Execute uma análise sobre a demanda de workload
- COST09-BP02 Implemente um buffer ou controle de utilização para gerenciar a demanda
- COST09-BP03 Forneça recursos dinamicamente

COST09-BP01 Execute uma análise sobre a demanda de workload

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 workload. O esforço de análise deve refletir o benefício potencial. Por exemplo, se o tempo gasto é proporcional ao custo da workload.

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

Orientação de implementação

conheça os requisitos da carga de trabalho. Os requisitos da organização devem indicar os tempos de resposta da carga de trabalho para solicitações. O tempo de resposta pode ser usado para determinar se a demanda é gerenciada ou se a oferta de recursos será alterada para atender à demanda.

A análise deve incluir a previsibilidade e a repetibilidade da demanda, a taxa de alteração na demanda e a quantidade de alteração na demanda. Verifique se a análise é realizada durante um

período longo o suficiente para incorporar qualquer variação sazonal, como processamento de fim de mês ou picos de fim de ano.

Certifique-se de que o esforço de análise reflita os possíveis benefícios da implementação da escalabilidade. Observe o custo total esperado do componente e quaisquer aumentos ou diminuições no uso e no custo durante a vida útil da carga de trabalho.

Você pode usar o AWS Cost Explorer ou Amazon QuickSight com o AWS Cost and Usage Report (CUR) ou seus logs de aplicação para fazer uma análise visual da demanda da workload.

Etapas da implementação

- Analisar dados de workload existentes: Analise dados da carga de trabalho existentes, das versões anteriores da carga de trabalho ou dos padrões de uso previstos. Use arquivos de log e dados de monitoramento para obter informações sobre como os clientes usam a carga de trabalho. As métricas típicas são a demanda real, em solicitações por segundo, os horários em que a taxa de demanda muda ou quando ela está em diferentes níveis e a taxa de alteração da demanda. Verifique se você analisou um ciclo completo da carga de trabalho, garantindo a coleta de dados para quaisquer alterações sazonais, como eventos de fim de mês ou de ano. O esforço refletido na análise deve refletir as características da carga de trabalho. O maior esforço deve ser colocado em workloads de alto valor com as maiores alterações na demanda. O menor esforço deve ser colocado em workloads de baixo valor que tenham alterações mínimas na demanda. Métricas comuns de valor são risco, reconhecimento da marca, receita ou custo da carga de trabalho.
- Prever a influência externa: Encontre membros da equipe de toda a organização que possam influenciar ou alterar a demanda na carga de trabalho. Equipes comuns são vendas, marketing ou desenvolvimento de negócios. Trabalhe com elas para saber os ciclos com os quais operam e se há eventos que alteram a demanda da workload. Preveja a demanda da carga de trabalho com esses dados.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Instance Scheduler
- Conceitos básicos do Amazon SQS
- AWS Cost Explorer
- Amazon QuickSight

COST09-BP02 Implemente 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. Verifique se os controles de utilização e buffers são projetados para que os clientes recebam uma resposta no tempo necessário.

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

Orientação de implementação

Controle de utilização: se a origem da demanda tiver capacidade de repetição, você poderá implementar o controle de utilização. O controle de utilização informa à origem que, se ela não puder atender à solicitação no momento atual, deverá tentar novamente mais tarde. A origem aguardará um período e, em seguida, tentará novamente a solicitação. A implementação do controle de utilização tem a vantagem de limitar a quantidade máxima de recursos e custos da carga de trabalho. Na AWS, você pode usar o Amazon API Gateway para implementar o controle de utilização. Consulte o whitepaper sobre o Pilar Confiabilidade do Well-Architected para obter mais detalhes sobre como implementar o controle de utilização.

Baseado em buffer: semelhante ao controle de utilização, um buffer adia o processamento de solicitações, permitindo que aplicativos executados em diferentes taxas se comuniquem com eficácia. Uma abordagem baseada em buffer usa uma fila para aceitar mensagens (unidades de trabalho) de produtores. As mensagens são lidas pelos consumidores e processadas, permitindo que as mensagens sejam executadas na taxa que atenda aos requisitos de negócios dos consumidores. Você não precisa se preocupar com os produtores que precisam lidar com problemas de controle de utilização, como durabilidade de dados e pressão contrária (onde os produtores ficam lentos porque o consumidor está correndo lentamente).

Na AWS, você pode escolher entre vários serviços para implementar uma abordagem de buffering. Amazon Simple Queue Service(Amazon SQS) é um serviço gerenciado que fornece filas que permitem que um único consumidor leia mensagens individuais. Amazon Kinesis oferece um fluxo que permite a muitos consumidores lerem as mesmas mensagens.

Ao criar uma arquitetura com uma abordagem baseada em buffer, certifique-se de arquitetar sua carga de trabalho para atender à solicitação no tempo necessário e de lidar com solicitações duplicadas de trabalho.

Etapas da implementação

- Analisar os requisitos do cliente: Analise as solicitações do cliente para determinar se são capazes de executar novas tentativas. Para clientes que não podem executar novas tentativas, buffers precisarão ser implementados. Analise a demanda geral, a taxa de alteração e o tempo de resposta necessário para determinar o tamanho do controle de utilização ou do buffer necessário.
- Implementar um buffer ou controle de utilização: Implemente um buffer ou um controle de utilização na carga de trabalho. Uma fila como Amazon Simple Queue Service (Amazon SQS) pode fornecer um buffer para seus componentes de workload. O Amazon API Gateway pode fornecer controle de utilização para seus componentes de workload.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Instance Scheduler
- Amazon API Gateway
- Amazon Simple Queue Service
- Conceitos básicos do Amazon SQS
- Amazon Kinesis

COST09-BP03 Forneça 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.

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

Orientação de implementação

Você pode usar o AWS Auto Scalingou incorporar escalabilidade em seu código com as API ou o SDK da AWS. Isso reduz os custos gerais da carga de trabalho removendo o custo operacional de fazer alterações manualmente em seu ambiente e pode ser executado muito mais rapidamente. Isso garantirá que o recurso da carga de trabalho corresponda melhor à demanda a qualquer momento.

Oferta baseada em demanda: aproveite a elasticidade da nuvem para fornecer recursos para atender à demanda em constante mudança. Aproveite as APIs ou os recursos de serviço para variar

programaticamente a quantidade de recursos de nuvem em sua arquitetura dinamicamente. Isso permite que você ajuste a escala de componentes em sua arquitetura e aumente automaticamente o número de recursos durante picos de demanda para manter a performance e reduzir a capacidade quando a demanda diminui para reduzir os custos.

AWS Auto Scaling ajuda você a ajustar sua capacidade para manter uma performance estável e previsível pelo menor custo possível. É um serviço totalmente gerenciado e gratuito que se integra a instâncias do Amazon Elastic Compute Cloud (Amazon EC2) e a frotas spot, ao Amazon Elastic Container Service (Amazon ECS), ao Amazon DynamoDB e ao Amazon Aurora.

O Auto Scaling oferece descoberta automática de recursos para ajudar a encontrar recursos na sua workload que possam ser configurados, tem estratégias de escalabilidade incorporadas para otimizar performance, custos ou um equilíbrio entre os dois, além de oferecer escalabilidade preditiva para ajudar com picos que ocorrem regularmente.

O Auto Scaling pode implementar escalabilidade manual, programada ou baseada em demanda. Você também pode usar métricas e alarmes de Amazon CloudWatch para acionar eventos de escalabilidade para sua carga de trabalho. As métricas típicas podem ser métricas padrão do Amazon EC2, como utilização de CPU, throughput de rede e latência de solicitação/resposta observada pelo Elastic Load Balancing(ELB) . Quando possível, você deve usar uma métrica que seja indicativa da experiência do cliente. Normalmente, essa é uma métrica personalizada que pode se originar do código da aplicação em sua workload.

Ao arquitetar com uma abordagem baseada em demanda, tenha em mente dois pontos essenciais. Primeiro, entenda a rapidez com que você deve provisionar novos recursos. Segundo, entenda que o tamanho da margem entre oferta e demanda mudará. Você deve estar pronto para lidar com a taxa de alteração na demanda e também estar pronto para falhas de recursos.

ELB ajuda a escalar distribuindo a demanda entre vários recursos. À medida que implementa mais recursos, você os adiciona ao balanceador de carga para atender à demanda. O Elastic Load Balancing tem suporte para instâncias do Amazon EC2, contêineres, endereços IP e funções do AWS Lambda.

Oferta baseada em tempo: Uma abordagem baseada em tempo alinha a capacidade de recurso a uma demanda que é previsível ou bem definida no tempo. Essa abordagem costuma não depender dos níveis de utilização dos recursos. Uma abordagem baseada em tempo garante que os recursos estejam disponíveis no momento específico em que são necessários e podem ser fornecidos sem nenhum atraso devido a procedimentos de inicialização e verificações do sistema ou de consistência.

Usando uma abordagem baseada em tempo, você pode fornecer recursos adicionais ou aumentar a capacidade durante períodos ocupados.

Você pode usar o Auto Scaling programado para implementar uma abordagem baseada em tempo. As cargas de trabalho podem ser programadas para expandir ou reduzir em horários definidos (por exemplo, o início do horário comercial), garantindo assim que os recursos estejam disponíveis quando os usuários ou a demanda chegarem.

Você também pode aproveitar as APIs e os SDKs da AWS e AWS CloudFormation para provisionar e desativar automaticamente ambientes inteiros conforme necessário. Essa abordagem é adequada para ambientes de desenvolvimento ou teste que são executados apenas nos períodos ou horários comerciais definidos.

Você pode usar APIs para ajustar a escala dos recursos dentro de um ambiente (ajuste de escala vertical). Por exemplo, você pode escalar uma carga de trabalho de produção alterando o tamanho ou a classe da instância. Isso pode ser feito interrompendo e iniciando a instância e selecionando a classe ou o tamanho da instância diferente. Essa técnica também pode ser aplicada a outros recursos, como Volumes elásticos do Amazon Elastic Block Store (Amazon EBS), que podem ser modificados para aumentar o tamanho, ajustar a performance (IOPS) ou alterar o tipo de volume durante o uso.

Ao arquitetar com uma abordagem baseada em tempo, tenha em mente dois pontos essenciais. Primeiro, qual é a consistência do padrão de uso? Segundo, qual será o impacto se o padrão mudar? Você pode aumentar a precisão das previsões monitorando suas cargas de trabalho e usando inteligência de negócios. Se você vir alterações significativas no padrão de uso, poderá ajustar os tempos para garantir que a cobertura seja fornecida.

Etapas da implementação

- Configure a programação baseada em tempo: Para alterações previsíveis na demanda, a escalabilidade baseada em tempo pode fornecer a quantidade correta de recursos em tempo hábil. Também será útil se a criação e a configuração de recursos não forem rápidas o suficiente para responder a alterações na demanda. Usando a análise de workload, configure a escalabilidade programada usando o AWS Auto Scaling.
- Configure o Auto Scaling: Para configurar a escalabilidade com base em métricas de carga de trabalho ativas, use o Amazon Auto Scaling. Use a análise e configure o Auto Scaling para acionar nos níveis de recursos corretos e garanta que a carga de trabalho seja dimensionada no tempo necessário.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Instance Scheduler
- Conceitos básicos do Amazon EC2 Auto Scaling
- Conceitos básicos do Amazon SQS
- Escalabilidade programada para o Amazon EC2 Auto Scaling

Otimizar ao longo do tempo

Pergunta

- COST 10 Como você avalia os novos serviços?

COST 10 Como você avalia os novos serviços?

À medida que a AWS lança novos serviços e recursos, uma das práticas recomendadas é avaliar suas decisões sobre a arquitetura existente a fim de garantir que elas ofereçam o melhor custo-benefício.

Práticas recomendadas

- COST10-BP01 Desenvolver um processo de análise da workload
- COST10-BP02 Revise e analisar a workload regularmente

COST10-BP01 Desenvolver um processo de análise da workload

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, workloads principais ou workloads com valor superior a 10% da fatura são analisadas trimestralmente, enquanto workloads abaixo de 10% são analisadas anualmente.

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

Orientação de implementação

Para garantir que você sempre tenha a workload mais econômica, é necessário revisar regularmente a workload para saber se há oportunidades de implementar novos serviços, recursos e componentes. Para garantir que você atinja custos gerais mais baixos, o processo deve ser proporcional à quantidade potencial de economia. Por exemplo, as cargas de trabalho que representam 50% do seu gasto geral devem ser analisadas com mais frequência e mais precisão do que as cargas de trabalho que representam 5% do seu gasto geral. Leve em consideração quaisquer fatores externos ou volatilidade. Se a carga de trabalho atender a uma área geográfica ou segmento de mercado específico e houver previsão de mudanças nessa área, revisões mais frequentes poderão resultar em economias de custos. Outro fator em análise é o esforço para implementar alterações. Se houver custos significativos em testes e validação de alterações, as revisões devem ser menos frequentes.

Leve em consideração o custo de longo prazo de manutenção de componentes e recursos obsoletos e na incapacidade de implementar novos recursos neles. O custo atual de testes e validação pode exceder o benefício proposto. No entanto, ao longo do tempo, o custo de fazer a mudança pode aumentar significativamente à medida que a lacuna entre a carga de trabalho e as tecnologias atuais aumenta, resultando em custos ainda maiores. Por exemplo, o custo da migração para uma nova linguagem de programação pode não ser econômico no momento. No entanto, em cinco anos, o custo de pessoas com qualificações nessa linguagem pode aumentar e, devido ao crescimento da carga de trabalho, você estaria movendo um sistema ainda maior para a nova linguagem, exigindo ainda mais esforço do que anteriormente.

Divida sua carga de trabalho em componentes, atribua o custo do componente (uma estimativa é suficiente) e liste os fatores (por exemplo, esforço e mercados externos) ao lado de cada componente. Use esses indicadores para determinar uma frequência de revisão para cada carga de trabalho. Por exemplo, você pode ter servidores web como um alto custo, baixo esforço de alteração e altos fatores externos, resultando em alta frequência de revisão. Um banco de dados central pode ser de custo médio, alto esforço de alteração e baixos fatores externos, resultando em uma média frequência de análise.

Etapas da implementação

- Definir frequência de análise: Defina a frequência com que a carga de trabalho e os componentes dela devem ser analisados. Essa é uma combinação de fatores, e pode diferir de carga de trabalho para carga de trabalho dentro da sua organização, além de também poder diferir entre componentes na carga de trabalho. Os fatores comuns incluem a importância para a organização medida em termos de receita ou marca, o custo total da execução da workload (incluindo custos

operacionais e de recursos), a complexidade da workload, a facilidade da implementação de uma alteração, qualquer contrato de licenciamento de software e se uma alteração incorreria em aumentos significativos nos custos de licenciamento devido a licenciamento punitivo. Os componentes podem ser definidos de maneira funcional ou técnica, como bancos de dados e servidores web ou recursos de computação e armazenamento. Equilibre os fatores de acordo e desenvolva um período para a carga de trabalho e os componentes dela. Você pode decidir revisar a workload completa a cada 18 meses, revisar os servidores Web a cada 6 meses, o banco de dados a cada 12 meses, computação e armazenamento de curto prazo a cada 6 meses e armazenamento de longo prazo a cada 12 meses.

- Definir a minuciosidade da análise: Defina quanto esforço é gasto na análise da carga de trabalho ou dos componentes da carga de trabalho. Semelhante à frequência da análise, esse é um equilíbrio de vários fatores. Você pode decidir gastar uma semana de análise no componente do banco de dados e quatro horas para análises de armazenamento.

Recursos

Documentos relacionados:

- Blog de novidades da AWS
- Tipos de computação em nuvem
- Quais as novidades da AWS

COST10-BP02 Revise e analisar a workload regularmente

As workloads existentes são analisadas regularmente de acordo com cada processo definido.

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

Orientação de implementação

Para obter os benefícios de novos serviços e recursos da AWS, você deve executar o processo de análise em suas workloads e implementar novos serviços e recursos, conforme necessário. Por exemplo, você pode revisar suas workloads e substituir o componente de mensagens por Amazon Simple Email Service (Amazon SES). Isso remove o custo de operação e manutenção de uma frota de instâncias e, ao mesmo tempo, fornece toda a funcionalidade a um custo reduzido.

Etapas da implementação

- Analise regularmente a workload: Usando o processo definido, execute análises com a frequência especificada. Verifique se você gastou a quantidade correta de esforço em cada componente. Esse processo seria semelhante ao processo de design inicial em que você selecionou serviços para otimização de custos. Analise os serviços e os benefícios que eles trariam, esse fator de tempo no custo de fazer a mudança, não apenas os benefícios de longo prazo.
- Implemente novos serviços: Se o resultado da análise for implementar alterações, primeiro execute uma linha de base da workload para saber o custo atual por cada saída. Implemente as alterações e, em seguida, execute uma análise para confirmar o novo custo por cada saída.

Recursos

Documentos relacionados:

- Blog de novidades da AWS
- Tipos de computação em nuvem
- Quais as novidades da AWS

Sustentabilidade

Tópicos

Escolha de região

Pergunta

SUS 1 Como você escolhe as regiões para apoiar suas metas de sustentabilidade?

Escolha as regiões onde você vai implementar suas workloads com base em seus requisitos empresariais e em suas metas de sustentabilidade.

Prática recomendada:

SUS01-BP01 Escolher regiões próximas aos projetos de energia renovável da Amazon e regiões onde a rede tem uma pegada de carbono publicada menor que em outros locais (ou regiões).

Escolha regiões próximas aos projetos de energia renovável da Amazon e regiões onde a grade de intensidade de carbono publicada esteja abaixo de outros locais (ou regiões).

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

Escolha regiões próximas aos projetos de energia renovável da Amazon e regiões onde a grade de intensidade de carbono publicada esteja abaixo de outros locais (ou regiões).

Recursos

Documentos relacionados:

- Amazon Around the Globe (Amazon em torno do globo)
- Metodologia de energia renovável
- What to Consider when Selecting a Region for your Workloads (O que considerar ao selecionar uma região para suas workloads)

Padrões de comportamento do usuário

Pergunta

- SUS 2 Como você pode utilizar favoravelmente os padrões de comportamento do usuário para apoiar suas metas de sustentabilidade?

SUS 2 Como você pode utilizar favoravelmente os padrões de comportamento do usuário para apoiar suas metas de sustentabilidade?

A maneira como os usuários consomem suas workloads e outros recursos pode ajudar você a identificar melhorias para atingir as metas de sustentabilidade. Escale a infraestrutura de tal

forma que ela sempre corresponda à carga de usuários e implante apenas os recursos mínimos necessários para oferecer suporte aos usuários. Alinhe os níveis de serviço às necessidades do cliente. Posicione os recursos de maneira a limitar a rede necessária para que eles sejam consumidos pelos usuários. Remova ativos que não sejam utilizados. Identifique ativos criados que não são utilizados e pare de gerá-los. Forneça aos membros da sua equipe dispositivos compatíveis com suas necessidades e com impacto de sustentabilidade reduzido.

Práticas recomendadas:

SUS02-BP01 Escalar a infraestrutura com a carga dos usuários

Identifique períodos de baixa utilização ou sem utilização e reduza a escala dos recursos verticalmente para eliminar a capacidade em excesso e melhorar a eficiência.

Antipadrões comuns:

- Você não dimensiona sua infraestrutura de acordo com a carga de usuários.
- Você dimensiona sua infraestrutura manualmente o tempo todo.
- Você deixa a capacidade aumentada após um evento de escalabilidade, em vez de reduzir novamente.

Benefícios do estabelecimento desta prática recomendada: A configuração e os testes da elasticidade da workload ajudam a reduzir o impacto ambiental da workload, economizar dinheiro e manter as referências da performance. Você pode aproveitar a elasticidade na nuvem para dimensionar automaticamente a capacidade durante e depois de picos de carga dos usuários para garantir que esteja usando apenas o número exato de recursos necessários para atender às necessidades dos clientes.

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

Orientação de implementação

- A elasticidade corresponde à oferta de recursos que você tem face à demanda por estes recursos. Instâncias, contêineres e funções oferecem mecanismos para elasticidade, seja em combinação com a escalabilidade automática ou como um recurso do serviço. Use elasticidade em sua arquitetura para garantir que a workload possa reduzir a escala verticalmente com rapidez e facilidade durante o período de baixa carga de usuários:
  - Uso Amazon EC2 Auto Scaling para verificar se você tem o número correto de instâncias do Amazon EC2 disponíveis para processar a carga da aplicação.

  - Uso Application Auto Scaling para dimensionar automaticamente os recursos para serviços individuais da AWS além do Amazon EC2, como funções do Lambda ou serviços do Amazon Elastic Container Service (Amazon ECS).
  - Uso o dimensionador automático de cluster do Kubernetes para dimensionar automaticamente os clusters do Kubernetes na AWS.
- Verifique se as métricas para aumentar ou reduzir a escala verticalmente são validadas em relação ao tipo de workload que está sendo implantada. Se você estiver implantando uma aplicação de transcodificação de vídeo, espera-se que a utilização da CPU seja de 100%, e essa não deve ser sua métrica principal. Você pode usar uma métrica personalizada (como utilização de memória) para a política de escalabilidade, se necessário. Para escolher as métricas certas, considere a seguinte orientação para o Amazon EC2:
  - A métrica deve ser uma métrica de utilização válida e descrever o quanto uma instância está ocupada.
  - O valor da métrica deve aumentar ou diminuir proporcionalmente com o número de instâncias no grupo do Auto Scaling.
- Uso a escalabilidade dinâmica em vez de escalabilidade manual para seu grupo do Auto Scaling. Também recomendamos que você use políticas de escalabilidade de monitoramento do objetivo em sua escalabilidade dinâmica.
- Verifique se as implantações de workload podem lidar com eventos de aumento e redução vertical da escala. Crie cenários de teste para eventos de redução da escala a fim de garantir que a carga de trabalho se comporte conforme o esperado. Você pode usar o histórico de atividades para testar e verificar uma atividade de escalabilidade para um grupo do Auto Scaling.
- Avalie sua workload com relação a padrões previsíveis e, ao antecipar alterações previstas e planejadas na demanda, escale proativamente. Uso escalabilidade preditiva com o Amazon EC2 Auto Scaling para eliminar a necessidade de superprovisionar capacidade.

Recursos

Documentos relacionados:

- Conceitos básicos do Amazon EC2 Auto Scaling
- Escalabilidade preditiva para o EC2 com Machine Learning
- Analisar o comportamento dos usuários usando o Amazon OpenSearch Service, o Amazon Data Firehose e o Kibana
- O que é o Amazon CloudWatch?

- O que é o AWS X-Ray?
- Logs de fluxo da VPC
- Monitorar a carga do banco de dados com o Performance Insights no Amazon RDS
- Introdução de suporte nativo para escalabilidade preditiva com o Amazon EC2 Auto Scaling
- Como criar uma política do Amazon EC2 Auto Scaling baseada em uma métrica de utilização de memória (Linux)
- Apresentando o Karpenter: um dimensionador automático de clusters do Kubernetes de código aberto e alta performance

Vídeos relacionados:

- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (CMP202-R1) (Computação melhor, mais rápida e mais barata: otimização de custos com o Amazon EC2)

Exemplos relacionados:

- Laboratório: Exemplos de grupos do Amazon EC2 Auto Scaling
- Laboratório: Implementação de escalabilidade automática com o Karpenter

SUS02-BP02 Alinhar os SLAs com as metas de sustentabilidade

Defina e atualize as metas dos Acordos de Nível de Serviço (SLAs), como períodos de disponibilidade ou de retenção de dados de modo a minimizar o número de recursos exigidos para comportar sua workload e, ao mesmo tempo, continuar atendendo aos requisitos empresariais.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Defina SLAs que apoiem suas metas de sustentabilidade e, ao mesmo tempo, atendam aos seus requisitos empresariais.
- Redefina os SLAs de tal modo que eles atendam aos requisitos empresariais, não que os exceda.
- Faça compensações que reduzam significativamente os impactos na sustentabilidade em troca de reduções aceitáveis em níveis de serviço.

- Use padrões de design que priorizem funções essenciais aos negócios e permita níveis de serviço mais baixos (como objetivos de tempo de resposta ou de tempo de recuperação) para funções não essenciais.

Recursos

Documentos relacionados:

- Acordos de Nível de Serviço (SLAs) da AWS
- Importance of Service Level Agreement for SaaS Providers

Vídeos relacionados:

- Building Sustainably on AWS (Criação de sustentabilidade na AWS)

SUS02-BP03 Interromper a criação e a manutenção de ativos não utilizados

Analise os ativos de aplicações (como relatórios pré-compilados, conjuntos de dados e imagens estáticas) e os padrões de acesso aos ativos para identificar redundâncias, subutilização e possíveis alvos de desativação. Consolide ativos gerados com conteúdo redundante (por exemplo, relatórios mensais com saídas e conjuntos de dados que se sobreponham ou sejam comuns) para remover os recursos consumidos quando há duplicação de saídas. Desative ativos não utilizados (por exemplo, imagens de produtos que não são mais vendidos) para liberar os recursos consumidos e reduzir o número de recursos usados para comportar a workload.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Gerencie ativos estáticos e remova ativos que não são mais necessários.
- Gerencie ativos gerados e interrompa a geração e remova ativos que não são mais necessários.
- Consolide ativos gerados sobrepostos para remover o processamento redundante.
- Instrua terceiros a interromper a produção e o armazenamento de ativos gerenciados em seu nome que não sejam mais necessários.
- Instrua terceiros a consolidar ativos redundantes produzidos em seu nome.

Recursos

Documentos relacionados:

- Optimizing your AWS Infrastructure for Sustainability, Part II: Storage (Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte II: Armazenamento)

Vídeos relacionados:

- Building Sustainably on AWS (Criação de sustentabilidade na AWS)

SUS02-BP04 Otimizar a localização geográfica de workloads para locais dos usuários

Analise os padrões de acesso à rede para identificar de onde seus clientes estão se conectando geograficamente. Selecione regiões e serviços que reduzam a distância que o tráfego de rede deve percorrer para reduzir o total de recursos de rede necessários para comportar a workload.

Antipadrões comuns:

- Selecione a região da workload com base em sua localização.

Benefícios do estabelecimento desta prática recomendada: Implantar uma workload perto dos clientes proporciona a latência mais baixa enquanto reduz a movimentação de dados pela rede e reduz o impacto ambiental.

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

Orientação de implementação

- Selecione as regiões para implantação da workload com base nos seguintes elementos fundamentais:
  - Sua meta de sustentabilidade: conforme explicado em Seleção de região.
  - A localização dos seus dados: para aplicações com uso intenso de dados (como big data e machine learning), o código da aplicação deve ser executado o mais perto possível dos dados.
  - A localização dos usuários: para aplicações voltadas ao usuário, escolha uma região próxima da base de clientes da workload.
  - Outras restrições: leve em conta restrições como segurança e conformidade, conforme explicado em O que considerar ao selecionar uma região para suas workloads.

- Uso zonas locais da AWS para executar workloads como renderização de vídeo e aplicações de área de trabalho virtual com uso intenso de gráficos. As zonas locais permitem que você se beneficie de ter recursos de computação e armazenamento mais próximos dos usuários finais.
- Use armazenamento em cache local ou soluções de armazenamento em cache da AWS para recursos usados com frequência a fim de aumentar a performance, reduzir a movimentação de dados e reduzir o impacto ambiental.
  - Uso Amazon CloudFront para armazenar conteúdo estático em cache, como imagens, scripts e vídeos, bem como conteúdo dinâmico, como APIs ou aplicações Web.
  - Uso Amazon ElastiCache para armazenar conteúdo em cache para aplicações Web.
  - Uso DynamoDB Accelerator para adicionar aceleração na memória às suas tabelas do DynamoDB.
- Use serviços que podem ajudar você a executar código mais perto dos usuários da workload:
  - Uso O Lambda@Edge para operações com uso computacional intenso que são executadas quando objetos não estão no cache.
  - Uso funções do Amazon CloudFront para casos de uso simples como solicitações HTTP(s) ou manipulações de resposta que podem ser executadas por funções de curta duração.
  - Uso AWS IoT Greengrass para executar computação local, mensagens e armazenamento de dados em cache para dispositivos conectados.
- Use o agrupamento de conexões para permitir a reutilização de conexões e reduzir os recursos necessários.
- Use datastores distribuídos que não dependem de conexões persistentes e atualizações síncronas para fins de consistência com o objetivo de atender a populações regionais.
- Substitua a capacidade de rede estática pré-provisionada por capacidade dinâmica compartilhada e divida o impacto sobre a sustentabilidade da capacidade de rede com outros assinantes.

Recursos

Documentos relacionados:

- Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte III: Redes
- Documentação do Amazon ElastiCache
- O que é o Amazon CloudFront?
- Principais recursos do Amazon CloudFront
- O Lambda@Edge

- Funções do CloudFront
- AWS IoT Greengrass

Vídeos relacionados:

- Building Sustainably on AWS (Criação de sustentabilidade na AWS)

Exemplos relacionados:

- Workshops de redes da AWS

SUS02-BP05 Otimizar os recursos dos membros da equipe para as atividades realizadas

Otimize os recursos fornecidos aos membros da equipe para minimizar o impacto sobre a sustentabilidade e, ao mesmo tempo, atender às suas necessidades. Por exemplo, realize operações complexas, como renderização e compilação, em desktops compartilhados na nuvem com alta utilização em vez de em sistemas de usuário único subutilizados com alto consumo de energia.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Provisione estações de trabalho e outros dispositivos para alinhar a maneira como eles são usados.
- Use desktops virtuais e a transmissão de aplicações para limitar os requisitos de upgrade e dispositivos.
- Migre para a nuvem as tarefas do processador e as com uso intenso de memória.
- Avalie o impacto de processos e sistemas no ciclo de vida de seus dispositivos e escolha soluções que minimizem o requisito de substituição de dispositivos e, ao mesmo tempo, atendam aos requisitos empresariais.
- Implemente o gerenciamento remoto de dispositivos para reduzir as viagens de negócios.

Recursos

Documentos relacionados:

- O que é o Amazon WorkSpaces?

- Documentação do Amazon AppStream 2.0
- NICE DCV
- AWS Systems Manager Fleet Manager

Vídeos relacionados:

- Building Sustainably on AWS (Criação de sustentabilidade na AWS)

Padrões de software e arquitetura

Pergunta

- SUS 3 Como você aproveita os padrões de software e arquitetura para apoiar suas metas de sustentabilidade?

SUS 3 Como você aproveita os padrões de software e arquitetura para apoiar suas metas de sustentabilidade?

Implemente padrões que suavizem os picos de carga e mantenham a alta utilização consistente de recursos implantados para minimizar os recursos consumidos. Os componentes podem ficar ociosos devido à falta de uso por conta das mudanças no comportamento do usuário ao longo do tempo. Revise os padrões e a arquitetura para consolidar os componentes subutilizados a fim de aumentar a utilização geral. Retire os componentes que não são mais necessários. Saiba qual é a performance dos componentes de sua workload e otimize os componentes que consomem a maioria dos recursos. Esteja ciente dos dispositivos que seus clientes usam para acessar seus serviços e implemente padrões a fim de minimizar a necessidade de upgrades de dispositivos.

Práticas recomendadas:

SUS03-BP01 Otimizar o software e a arquitetura para trabalhos assíncronos e programados

Use designs e arquiteturas eficientes de software para minimizar a média de recursos necessários por unidade de trabalho. Implemente mecanismos que resultem em uma utilização uniforme de componentes para reduzir os recursos ociosos entre as tarefas e minimizar o impacto de picos de carga.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Enfileire solicitações que não exigem processamento imediato.
- Aumente a serialização para nivelar a utilização em todo o pipeline.
- Modifique a capacidade de componentes individuais para evitar que os recursos fiquem ociosos aguardando a entrada.
- Crie buffers e estabeleça limites de taxa para regular o consumo de serviços externos.
- Use o hardware mais eficiente disponível para suas otimizações de software.
- Use arquiteturas orientadas a filas, gerenciamento de pipelines e operadores de instância sob demanda para maximizar a utilização do processamento em lote.
- Programe tarefas para evitar os picos de carga e a contenção de recursos de execução simultânea.
- Programe trabalhos em horários do dia em que a intensidade de carbono para a geração de energia é menor.

Recursos

Documentos relacionados:

- O que é o Amazon Simple Queue Service?
- O que é o Amazon MQ?
- Escalabilidade baseada no Amazon SQS
- O que é o AWS Step Functions?
- O que é o AWS Lambda?
- Usar o AWS Lambda com o Amazon SQS
- O que é o Amazon EventBridge?

Vídeos relacionados:

- Building Sustainably on AWS (Criação de sustentabilidade na AWS)
- Moving to event-driven architectures (Mudar para arquiteturas orientadas a eventos)

SUS03-BP02 Remover ou refatorar componentes da workload subutilizados ou não utilizados

Monitore a atividade da workload para identificar alterações na utilização de componentes individuais ao longo do tempo. Remova os componentes que não são mais utilizados nem necessários e refatore os componentes subutilizados para limitar o desperdício de recursos.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Analise a carga (usando indicadores como fluxo de transações e chamadas de API) em componentes funcionais para identificar componentes não utilizados e subutilizados.
- Retire os componentes que não são mais necessários.
- Refatore os componentes subutilizados.
- Consolide os componentes subutilizados com outros recursos para melhorar a eficiência da utilização.

Recursos

Documentos relacionados:

- O que é o AWS X-Ray?
- O que é o Amazon CloudWatch?
- Usar o ServiceLens para monitorar a integridade das suas aplicações
- Automated Cleanup of Unused Images in Amazon ECR (Limpeza automatizada de imagens não utilizadas no Amazon ECR)

Vídeos relacionados:

- Building Sustainably on AWS (Criação de sustentabilidade na AWS)

SUS03-BP03 Otimizar as áreas de código que consomem mais tempo ou recursos

Monitore a atividade da workload para identificar os componentes da aplicação que consomem a maioria dos recursos. Otimize o código que é executado nesses componentes para minimizar o uso de recursos e, ao mesmo tempo, maximizar a performance.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Monitore a performance como uma função do uso de recurso para identificar componentes com requisitos de recursos altos por unidade de trabalho como alvos de otimização.
- Use um criador de perfil de código para identificar as áreas de código que gastam mais tempo ou usam mais recursos e as defina como alvos de otimização.
- Substitua algoritmos por versões mais eficientes que produzem o mesmo resultado.
- Use a aceleração de hardware para melhorar a eficiência de blocos de código com longos tempos de execução.
- Use a linguagem de programação e o sistema operacional mais eficientes para a workload.
- Remova classificações e formatações desnecessárias.
- Use padrões de transferência de dados que minimizem os recursos com base na frequência de alterações dos dados e em como eles são consumidos. Por exemplo, envie informações sobre alterações de estado para um cliente em vez de fazê-lo consumir recursos para executar sondagens e receber mensagens inúteis que relatem “não há alteração”.

Recursos

Documentos relacionados:

- O que é o Amazon CloudWatch?
- O que é o Amazon CodeGuru Profiler?
- Instâncias de FPGA
- Os AWS SDKs em Ferramentas para desenvolver na AWS

Vídeos relacionados:

- Building Sustainably on AWS (Criação de sustentabilidade na AWS)

SUS03-BP04 Otimizar o impacto sobre dispositivos e equipamentos de clientes

Conheça os dispositivos e o equipamento que os clientes usam para consumir seus serviços, o ciclo de vida esperado para eles e o impacto financeiro e na sustentabilidade pela substituição desses

componentes. Implemente padrões e arquiteturas de software de modo a minimizar a necessidade de substituir dispositivos e fazer upgrade de equipamento. Por exemplo, implemente novos recursos usando código compatível com versões anteriores de sistemas operacionais e hardware mais antigos, ou gerencie o tamanho das cargas úteis para que elas não excedam a capacidade de armazenamento do dispositivo de destino.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Faça um inventário dos dispositivos usados pelos clientes.
- Teste usando farms de dispositivos gerenciados com conjuntos representativos de hardware para entender o impacto de suas alterações, e itere o desenvolvimento para maximizar os dispositivos compatíveis.
- Considere a largura de banda da rede e a latência ao criar cargas úteis, e implemente recursos que ajudem suas aplicações a funcionar bem em links de baixa largura de banda e alta latência.
- Pré-processe cargas úteis de dados para reduzir os requisitos de processamento local e limitar os requisitos de transferência de dados.
- Realize atividades com computação intensa no lado do servidor (como renderização de imagens) ou use a transmissão de aplicações para melhorar a experiência do usuário em dispositivos mais antigos.
- Faça a segmentação e a paginação dos dados de saída, especialmente para sessões interativas, a fim de gerenciar cargas úteis e limitar os requisitos de armazenamento local.

Recursos

Documentos relacionados:

- O que é o AWS Device Farm?
- Documentação do Amazon AppStream 2.0
- NICE DCV
- Documentação do Amazon Elastic Transcoder

Vídeos relacionados:

- Building Sustainably on AWS (Criação de sustentabilidade na AWS)

SUS03-BP05 Usar arquiteturas e padrões de software que oferecem suporte melhor aos padrões de armazenamento e ao acesso a dados

Entenda como os dados são usados com sua workload, consumidos pelos usuários, transferidos e armazenados. Escolha tecnologias com o mínimo de requisitos de armazenamento e processamento de dados.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Analise seus padrões de armazenamento e acesso a dados.
- Armazene arquivos de dados em formatos de arquivo eficientes, como Parquet, para evitar processamento desnecessário (por exemplo, ao executar análises) e para reduzir o armazenamento total provisionado.
- Use tecnologias que funcionam nativamente com dados compactados.
- Use o mecanismo de banco de dados que melhor comporta seu padrão de consulta dominante.
- Gerencie seus índices de bancos de dados para garantir que o design do índice comporte a execução eficiente de consultas.
- Escolha protocolos de rede que reduzam a quantidade de capacidade de rede consumida.

Recursos

Documentos relacionados:

- Formatos de arquivos compactados compatíveis com o Athena
- COPY de formatos de dados colunares com o Amazon Redshift
- Converter o formato de registro de entrada no Firehose
- Opções de formato para entradas e saídas de ETL no AWS Glue
- Melhorar a performance de consultas no Amazon Athena fazendo a conversão para formatos colunares
- Carregar arquivos de dados compactados do Amazon S3 com o Amazon Redshift
- Monitorar a carga do banco de dados com o Performance Insights no Amazon Aurora
- Monitorar a carga do banco de dados com o Performance Insights no Amazon RDS
- AWS IoT FleetWise

Vídeos relacionados:

- Building Sustainably on AWS (Criação de sustentabilidade na AWS)

Padrões de dados

Pergunta

- SUS 4 Como você aproveita o acesso a dados e os padrões de uso para apoiar suas metas de sustentabilidade?

SUS 4 Como você aproveita o acesso a dados e os padrões de uso para apoiar suas metas de sustentabilidade?

Implemente práticas de gerenciamento de dados para reduzir o armazenamento provisionado necessário para comportar a workload e os recursos exigidos para usá-la. Entenda seus dados e use as tecnologias e as configurações de armazenamento que melhor promovam o valor empresarial dos dados e a forma como eles são usados. Gerencie o ciclo de vida dos dados e opte por um armazenamento mais eficiente e com menor performance quando os requisitos diminuírem, excluindo os dados que não são mais necessários.

Práticas recomendadas:

SUS04-BP01 Implementar uma política de classificação de dados

Classifique dados para entender seu significado para os resultados dos negócios. Use essas informações para determinar quando é possível migrar os dados para um armazenamento com uso mais eficiente de energia ou excluí-los de forma segura.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Determine os requisitos de distribuição, retenção e exclusão dos dados.
- Use a marcação em volumes e objetos para registrar os metadados que são usados para determinar como eles devem ser gerenciados, incluindo a classificação dos dados.
- Audite periodicamente seu ambiente em busca de dados que não estejam etiquetados ou classificados e classifique-os e etiquete-os apropriadamente.

Recursos

Documentos relacionados:

- Processo de classificação de dados
- Utilização da Nuvem AWS para oferecer suporte à classificação de dados
- Políticas de tags do AWS Organizations

SUS04-BP02 Usar tecnologias compatíveis com seus padrões de acesso e de armazenamento de dados

Use um armazenamento mais adequado à maneira como seus dados são acessados e armazenados a fim de reduzir os recursos provisionados e, ao mesmo tempo, comportar sua workload. Por exemplo, dispositivos de estado sólido (SSDs) consomem mais energia do que unidades magnéticas e devem ser usados somente para casos de uso de dados ativos. Use um armazenamento de classe de arquivamento com eficiência de energia para dados acessados com pouca frequência.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Monitore seus padrões de acesso a dados.
- Migre dados para a tecnologia apropriada com base no padrão de acesso.
- Migre dados de arquivamento para um armazenamento projetado para essa finalidade.

Recursos

Documentos relacionados:

- Tipos de volume do Amazon EBS
- Armazenamento de instâncias do Amazon EC2
- Amazon S3 Intelligent-Tiering
- Uso de classes de armazenamento do Amazon S3
- O que é o Amazon CloudWatch?
- O que é o Amazon S3 Glacier?

Vídeos relacionados:

- Architectural Patterns for Data Lakes on AWS (Padrões de arquitetura para data lakes na AWS)

SUS04-BP03 Usar políticas de ciclo de vida para excluir dados desnecessários

Gerencie o ciclo de vida de todos os seus dados e defina cronogramas de exclusão automática para minimizar os requisitos totais de armazenamento da workload.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Defina políticas de ciclo de vida para todos os seus tipos de classificação de dados.
- Defina políticas automatizadas de ciclo de vida para aplicar regras de ciclo de vida.
- Exclua volumes e snapshots não utilizados.
- Agregue dados quando aplicável com base nas regras de ciclo de vida.

Recursos

Documentos relacionados:

- Políticas de ciclo de vida do Amazon ECR
- Gerenciamento de ciclo de vida do Amazon EFS
- Amazon S3 Intelligent-Tiering
- Avaliar recursos com o Regras do AWS Config
- Gerenciamento do ciclo de vida de armazenamento no Amazon S3
- Políticas de ciclo de vida de objetos no AWS Elemental MediaStore

Vídeos relacionados:

- Amazon S3 Lifecycle (Ciclo de vida do Amazon S3)

SUS04-BP04 Minimizar o provisionamento em excesso no armazenamento em bloco

Para reduzir o armazenamento total provisionado, crie um armazenamento em bloco com alocações por tamanho que sejam apropriadas para a workload. Use volumes elásticos para expandir o

armazenamento à medida que os dados aumentam sem precisar redimensionar o armazenamento anexado aos recursos de computação. Analise regularmente volumes elásticos e reduza volumes com excesso de provisionamento para se ajustar ao tamanho de dados atual.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Monitore a utilização dos seus volumes de dados.
- Use volumes elásticos e serviços gerenciados de dados em bloco para automatizar a alocação de armazenamento adicional à medida que os seus dados persistentes aumentarem.
- Defina os níveis pretendidos de utilização para seus volumes de dados e redimensione os volumes fora dos intervalos esperados.
- Dimensione volumes somente leitura para acomodar os dados.
- Migre os dados para depósitos de objetos a fim de evitar o provisionamento de capacidade em excesso que ocorre com os tamanhos de volumes fixos no armazenamento em bloco.

Recursos

Documentos relacionados:

- Volumes elásticos do Amazon EBS
- Documentação do Amazon FSx
- O que é o Amazon CloudWatch?
- O que é o Amazon Elastic File System?

SUS04-BP05 Remover dados desnecessários ou redundantes

Duplique os dados somente quando necessário para reduzir o armazenamento total consumido. Use tecnologias de backup que eliminem dados duplicados em níveis de arquivo e bloco. Limite o uso de configurações RAID (Matriz redundante de unidades independentes), exceto quando necessário para atender aos acordos de nível de serviço (SLAs).

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Use mecanismos que possam duplicar dados no nível de bloco e objeto.

- Use tecnologias de backup que possam fazer backups incrementais e eliminação da duplicação de dados em níveis de bloco, arquivo e objeto.
- Use o RAID somente quando for necessário para atender aos SLAs.
- Centralize o log e rastreie os dados, elimine a duplicação de entradas de log idênticas e estabeleça mecanismos para ajustar a prolixidade quando necessário.
- Preencha os caches com antecedência somente quando justificável.
- Estabeleça o monitoramento e a automação de cache para redimensioná-lo de forma adequada.
- Remova implantações e ativos desatualizados de depósitos de objetos e caches de borda ao enviar novas versões da sua workload.

Recursos

Documentos relacionados:

- Snapshots do Amazon EBS
- Retenção de dados do log de alterações no CloudWatch Logs
- Eliminação de duplicação de dados no Amazon FSx for Windows File Server
- Recursos do Amazon FSx for ONTAP incluindo a eliminação da duplicação de dados
- Invalidar arquivos no Amazon CloudFront
- Usar o AWS Backup para fazer backup e restaurar sistemas de arquivos do Amazon EFS
- O que é o Amazon CloudWatch Logs?
- Trabalhar com backups no Amazon RDS

Exemplos relacionados:

- Laboratório: Optimize Data Pattern using Amazon Redshift Data Sharing (Otimizar o padrão de dados usando o compartilhamento de dados do Amazon Redshift)

SUS04-BP06 Usar sistemas de arquivos compartilhados ou armazenamento de objetos para acessar dados comuns

Adote o armazenamento compartilhado e fontes únicas de verdade para evitar duplicação de dados e reduzir os requisitos totais de armazenamento da workload. Busque dados do armazenamento compartilhado somente conforme necessário. Desanexe volumes não utilizados para disponibilizar mais recursos.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Migre dados para o armazenamento compartilhado quando eles tiverem vários consumidores.
- Busque dados do armazenamento compartilhado somente conforme necessário.
- Exclua os dados conforme apropriado para seus padrões de uso e implemente a funcionalidade de tempo de vida (TTL) para gerenciar dados armazenados em cache.
- Desvincule volumes de clientes que não estão utilizando-os ativamente.

Recursos

Documentos relacionados:

- Amazon FSx
- Estratégias de armazenamento em cache
- O que é o Amazon Elastic File System?
- O que é o Amazon S3?

SUS04-BP07 Minimizar a movimentação de dados entre redes

Use o armazenamento compartilhado e acesse dados de datastores regionais para minimizar os recursos totais de rede exigidos para comportar a movimentação de dados da workload.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Armazene dados o mais próximo possível do consumidor.
- Particione serviços consumidos regionalmente para que os dados específicos da região sejam armazenados na região em que eles são consumidos.
- Use a duplicação em nível de bloco em vez da duplicação em nível de arquivo ou objeto ao copiar alterações na rede.
- Compacte os dados antes de transferi-los pela rede.

Recursos

Documentos relacionados:

- Optimizing your AWS Infrastructure for Sustainability, Part III: Networking (Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte III: Redes)
- Infraestrutura global da AWS
- Principais recursos do Amazon CloudFront incluindo a rede global de borda do CloudFront
- Compactação de solicitações HTTP no Amazon OpenSearch Service
- Intermediar a compactação de dados com o Amazon EMR
- Carregar arquivos de dados compactados do Amazon S3 no Amazon Redshift
- Distribuição de arquivos compactados com o Amazon CloudFront

SUS04-BP08 Fazer backup de dados somente quando for difícil recriar

Para reduzir o consumo de armazenamento, faça backup somente de dados com valor empresarial ou que sejam necessários para atender aos requisitos de conformidade. Examine as políticas de backup e exclua armazenamentos temporários que não forneçam valor em um cenário de recuperação.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Baixo

Orientações para a implementação

- Use sua classificação de dados para estabelecer de quais dados é necessário fazer backup.
- Exclua dados que você possa recriar facilmente.
- Exclua dados temporários dos seus backups.
- Exclua cópias locais de dados, a menos que o tempo necessário para restaurar esses dados de um local comum exceda seus Acordos de Serviço (SLAs).

Recursos

Documentos relacionados:

- Usar o AWS Backup para fazer backup e restaurar sistemas de arquivos do Amazon EFS
- Snapshots do Amazon EBS
- Trabalhar com backups no Amazon Relational Database Service

Padrões de hardware

Pergunta

- SUS 5 Como suas práticas de gerenciamento de hardware e de uso apoiam suas metas de sustentabilidade?

SUS 5 Como suas práticas de gerenciamento de hardware e de uso apoiam suas metas de sustentabilidade?

Procure oportunidades para reduzir os impactos na sustentabilidade da workload fazendo mudanças nas suas práticas de gerenciamento de hardware. Minimize a quantidade de hardware necessária para provisionar e implantar e escolha o hardware mais eficiente para sua workload individual.

Práticas recomendadas:

SUS05-BP01 Usar a quantidade mínima de hardware para atender às suas necessidades

Ao usar os recursos da nuvem, é possível fazer alterações frequentes às implementações da workload. Atualize os componentes implantados conforme suas necessidades mudarem.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a implementação

- Permita a escalabilidade horizontal e use a automação para aumentar a escala na horizontal à medida que as cargas aumentam e para reduzir a escala na horizontal à medida que as cargas diminuem.
- Escale usando pequenos incrementos para workloads variáveis.
- Alinhe a escalabilidade com os padrões de utilização cíclica (por exemplo, um sistema de folha de pagamento com atividades bissemanais de processamento intensas) à medida que a carga varia ao longo de dias, semanas, meses ou anos.
- Negocie Acordos de Nível de Serviço (SLAs) que permitam uma redução temporária na capacidade enquanto a automação implanta recursos de substituição.

Recursos

Documentos relacionados:

- Documentação do AWS Compute Optimizer

- Otimização do Lambda: otimização da performance
- Documentação do Auto Scaling

SUS05-BP02 Usar tipos de instância com o mínimo de impacto

Monitore continuamente o lançamento de novos tipos de instância e aproveite as melhorias de eficiência de energia, incluindo os tipos de instância projetados para comportar workloads específicas, como treinamento de machine learning, inferência e transcodificação de vídeo.

Antipadrões comuns:

- Você usa apenas uma família de instâncias.
- Você usa apenas instâncias x86.
- Você especifica um tipo de instância em sua configuração do Amazon EC2 Auto Scaling.
- Você usa instâncias da AWS de um modo para o qual elas não foram projetadas (por exemplo, você usa instâncias otimizadas para computação em uma workload com uso intenso de memória).
- Você não avalia os novos tipos de instância regularmente.
- Você não verifica as recomendações de ferramentas de dimensionamento correta da AWS, como o AWS Compute Optimizer.

Benefícios do estabelecimento desta prática recomendada: Ao usar instâncias com eficiência de energia e dimensionadas corretamente, você consegue reduzir ainda mais o impacto ambiental e o custo da workload.

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

Orientação de implementação

- Conheça e explore os tipos de instância que podem reduzir o impacto ambiental de sua workload.
  - Inscreva-se nas Novidades da AWS para ficar por dentro das tecnologias e instâncias mais recentes da AWS.
  - Conheça os diversos tipos de instâncias da AWS.
  - Conheça as instâncias baseadas em AWS Graviton, que oferecem a melhor performance por watt de energia usada no Amazon EC2 assistindo aos vídeos re:Invent 2020 - Deep dive on AWS Graviton2 processor-powered Amazon EC2 instances (re:Invent 2020 - aprofundamento em instâncias do Amazon EC2 alimentadas por processadores AWS Graviton2) e Deep dive

into AWS Graviton3 and Amazon EC2 C7g instances (Aprofundamento em AWS Graviton3 e instâncias C7g do Amazon EC2).

- Planeje e migre sua workload para tipos de instância com impacto mínimo.
  - Defina um processo para avaliar novos recursos ou instâncias para sua workload. Aproveite a agilidade da nuvem para testar rapidamente como novos tipos de instância podem melhorar a sustentabilidade ambiental de sua workload. Use métricas de proxy para mensurar quantos recursos são necessários para concluir uma unidade de trabalho.
  - Se possível, modifique sua workload para trabalhar com diferentes números de vCPUs e diferentes quantidades de memória para maximizar sua escolha de tipo de instância.
  - Considere migrar sua workload para instâncias baseadas em Graviton e melhorar a eficiência da performance da workload (consulte AWS Graviton Fast Start e AWS Graviton2 para ISVs). Tenha em mente as considerações ao migrar workloads para instâncias do Amazon Elastic Compute Cloud baseadas em AWS Graviton.
  - Considere selecionar a opção AWS Graviton em seu uso de serviços gerenciados da AWS.
  - Migre sua workload para regiões que ofereçam instâncias com o menor impacto na sustentabilidade e atendam aos seus requisitos de negócios.
  - Para workloads de machine learning, use instâncias do Amazon EC2 que se baseiam em chips personalizados do Amazon Machine Learning como AWS Trainium, AWS Inferentiae aos Amazon EC2 DL1.
  - Uso Amazon SageMaker Inference Recommender para dimensionar endpoints de inferência de ML corretamente.
  - Para workloads com transcodificação de vídeo em tempo real, use instâncias VT1 do Amazon EC2.
  - Para workloads com picos (workloads com requisitos infrequentes para capacidade adicional), use instâncias de performance expansível.
  - Para workloads sem estado e tolerantes a falhas, use Instâncias Spot do Amazon EC2 para aumentar a utilização geral da nuvem e reduzir o impacto na sustentabilidade de recursos não utilizados.
- Opere e otimize a instância de sua workload.
  - Para workloads efêmeras, avalie métricas do Amazon CloudWatch para instâncias , como `CPUUtilization` , a fim de identificar se a instância está ociosa ou é subutilizada.
  - Para workloads estáveis, verifique as ferramentas da AWS para dimensionamento correto, como o AWS Compute Optimizer , em intervalos regulares a fim de identificar oportunidades para otimizar e dimensionar instâncias corretamente.

Recursos

Documentos relacionados:

- Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte I: Computação
- Processador AWS Graviton
- AWS Inferentia
- AWS Trainium
- Amazon EC2 DL1
- Instâncias de performance expansível do Amazon EC2
- Frotas de reserva de capacidade do Amazon EC2
- Frota spot do Amazon EC2
- Instâncias Spot do Amazon EC2
- Instâncias VT1 do Amazon EC2
- Tipos de instância do Amazon EC2
- AWS Compute Optimizer
- Funções: configuração de função do Lambda

Vídeos relacionados:

- Deep dive on AWS Graviton2 processor-powered Amazon EC2 instances (Aprofundamento em instâncias do Amazon EC2 alimentadas por processadores AWS Graviton2)
- Deep dive into AWS Graviton3 and Amazon EC2 C7g instances (Aprofundamento em AWS Graviton3 e instâncias C7g do Amazon EC2)

Exemplos relacionados:

- Laboratório: Recomendações de dimensionamento correto
- Laboratório: Dimensionamento correto com o Compute Optimizer
- Laboratório: Otimizar padrões de hardware e observar KPIs de sustentabilidade

SUS05-BP03 Usar serviços gerenciados

Os serviços gerenciados transferem para a AWS a responsabilidade pela manutenção de uma média elevada de utilização e pela otimização da sustentabilidade do hardware implantado na AWS. Use serviços gerenciados para distribuir o impacto na sustentabilidade do serviço entre todos os locatários dele, reduzindo sua contribuição individual.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Migre de serviços auto-hospedados para serviços gerenciados. Por exemplo, use as instâncias gerenciadas do Amazon Relational Database Service (Amazon RDS) , em vez de manter suas próprias instâncias do Amazon RDS no Amazon Elastic Compute Cloud (Amazon EC2)ou use serviços de contêiner, como o AWS Fargate, em vez de implementar sua própria infraestrutura de contêiner.

Recursos

Documentos relacionados:

- AWS Fargate
- Amazon DocumentDB
- Amazon Elastic Kubernetes Service (EKS)
- Amazon Managed Streaming for Apache Kafka (Amazon MSK)
- Amazon Redshift
- O Amazon Relational Database Service (RDS)

SUS05-BP04 Otimizar o uso de GPUs

Unidades de processamento gráfico (GPUs) podem ser uma fonte de alto consumo de energia, e várias workloads de GPU são altamente variáveis, como renderização, transcodificação e treinamento e modelagem de machine learning. Execute instâncias de GPU somente pelo tempo necessário e desative-as com automação quando não precisar mais delas para reduzir o consumo de recursos.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Use GPUs somente para tarefas em que elas sejam mais eficientes do que alternativas baseadas em CPU.
- Use automação para liberar instâncias de GPU quando não estiverem em uso.
- Use aceleração de gráficos flexível em vez de instâncias de GPU dedicadas.
- Aproveite o hardware personalizado específico para sua workload.

Recursos

Documentos relacionados:

- Computação acelerada
- AWS Inferentia
- AWS Trainium
- Computação acelerada para instâncias do EC2
- Instâncias do VT1 do Amazon EC2
- Amazon Elastic Graphics

Processo de desenvolvimento e implantação

Pergunta

- SUS 6 Como seus processos de desenvolvimento e implantação apoiam suas metas de sustentabilidade?

SUS 6 Como seus processos de desenvolvimento e implantação apoiam suas metas de sustentabilidade?

Procure oportunidades para reduzir seu impacto na sustentabilidade fazendo mudanças nas suas práticas de desenvolvimento, teste e implantação.

Práticas recomendadas:

SUS06-BP01 Adotar métodos que podem apresentar melhorias na sustentabilidade rapidamente

Teste e valide melhorias potenciais antes de implantá-las em produção. Considere o custo do teste ao calcular o benefício futuro potencial de uma melhoria. Desenvolva métodos de teste de baixo custo para permitir pequenas melhorias.

Nível de exposição a riscos quando esta prática recomendada não for estabelecida: Médio

Orientações para a implementação

- Adicione requisitos de sustentabilidade ao seu processo de desenvolvimento.
- Permita que os recursos trabalhem em paralelo com o objetivo de desenvolver, testar e implantar melhorias na sustentabilidade.
- Teste e valide as possíveis melhorias no impacto sobre a sustentabilidade antes de implantá-las na produção.
- Teste possíveis melhorias usando os componentes representativos mínimos viáveis.
- Implante melhorias na sustentabilidade testadas na produção à medida que elas forem disponibilizadas.

Recursos

Documentos relacionados:

- A AWS viabiliza soluções de sustentabilidade

Exemplos relacionados:

- Laboratório: Transformar relatórios de custo e uso em relatórios de eficiência

SUS06-BP02 Manter a workload atualizada

Sistemas operacionais, bibliotecas e aplicações atualizados podem melhorar a eficiência da workload e facilitar a adoção de tecnologias mais eficientes. Um software atualizado também pode incluir recursos para medir o impacto na sustentabilidade da workload com mais precisão, pois os fornecedores oferecem recursos para atender às suas próprias metas de sustentabilidade.

Antipadrões comuns:

- Você pressupõe que sua arquitetura atual se tornará estática sem nenhuma atualização ao longo do tempo.
- Você não tem nenhum sistema ou ritmo regular para avaliar se software ou pacotes atualizados são compatíveis com sua workload.
- Você apresenta alterações de arquitetura ao longo do tempo sem justificativa.

Benefícios do estabelecimento desta prática recomendada: Ao estabelecer um processo para manter a workload atualizada, você poderá adotar novos recursos e capacidades, resolver problemas e aumentar a eficiência da workload.

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

Orientação de implementação

- Defina um processo e um cronograma para avaliar novos recursos ou instâncias para sua workload. Aproveite a agilidade da nuvem para testar rapidamente como novos recursos podem melhorar sua workload com o intuito de:
  - Reduzir impactos de sustentabilidade.
  - Obter eficiências de performance.
  - Remover barreiras de melhorias planejadas.
  - Aumentar sua capacidade de medir e gerenciar impactos na sustentabilidade.
- Fazer o inventário de software e arquitetura da workload e identificar os componentes que precisam ser atualizados. Você pode usar o inventário do AWS Systems Manager para coletar metadados de sistema operacional (SO), aplicação e instância das instâncias do Amazon EC2 e entender rapidamente quais instâncias executam o software e as configurações exigidas pela política de software e quais instâncias precisam ser atualizadas.
- Entenda como atualizar os componentes de sua workload.
  - Gerencie atualizações para Amazon Machine Images (AMI) para imagens de servidor Linux ou Windows usando o EC2 Image Builder.
  - Use o Amazon Elastic Container Registry (Amazon ECR) com seu pipeline existente para gerenciar imagens do Amazon Elastic Container Service (Amazon ECS) e gerenciar imagens do Amazon Elastic Kubernetes Service.
  - O AWS Lambda inclui recursos de gerenciamento de versão.
- Use automação no processo de atualização para reduzir o nível de esforço para implantar novos recursos e limitar erros causados por processos manuais. Use ferramentas como o AWS Systems

Manager Patch Manager para automatizar o processo de atualizações do sistema e programar a atividade usando Janelas de Manutenção do AWS Systems Manager.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- Novidades da AWS
- Ferramentas de desenvolvedor da AWS
- AWS Systems Manager Patch Manager

Exemplos relacionados:

- Laboratórios do Well-Architected: Gerenciamento de inventário e patches
- Laboratório: AWS Systems Manager

SUS06-BP03 Aumentar a utilização de ambientes de desenvolvimento

Use a automação e a infraestrutura como código para ativar ambientes de pré-produção quando necessário e desativá-los quando não forem usados. Um padrão comum é programar períodos de disponibilidade que coincidam com as horas de trabalho dos membros da equipe de desenvolvimento. A hibernação é uma ferramenta útil para preservar o estado e colocar rapidamente as instâncias online apenas quando necessário. Use tipos de instância com capacidade de expansão, instâncias spot, serviços de banco de dados elásticos, contêineres e outras tecnologias para alinhar a capacidade de desenvolvimento e teste com o uso.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

- Use a automação para maximizar a utilização dos seus ambientes de desenvolvimento e teste.
- Use a automação para gerenciar o ciclo de vida dos seus ambientes de desenvolvimento e teste.
- Use ambientes representativos mínimos viáveis para desenvolver e testar possíveis melhorias.
- Use instâncias sob demanda para complementar os dispositivos de desenvolvedor.
- Use a automação para maximizar a eficiência dos seus recursos de compilação.

- Use tipos de instância com capacidade de expansão, instâncias spot e outras tecnologias para alinhar a capacidade de compilação com o uso.
- Adote serviços de nuvem nativos para acesso seguro ao shell de instância em vez de implantar frotas de hosts bastion.

Recursos

Documentos relacionados:

- Gerenciador de sessões do AWS Systems Manager
- Instâncias de performance expansível do Amazon EC2
- O que é o AWS CloudFormation?

SUS06-BP04 Usar farms de dispositivos gerenciados para testes

Farms de dispositivos gerenciados distribuem o impacto na sustentabilidade da fabricação do hardware e do uso de recursos entre vários locatários. Farms de dispositivos gerenciados oferecem diversos tipos de dispositivos para que você ofereça compatibilidade com componentes de hardware mais antigos e menos populares e evite o impacto sobre a sustentabilidade do cliente devido a atualizações desnecessárias de dispositivos.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: Baixo

Orientações para a implementação

Teste usando farms de dispositivos gerenciados com conjuntos representativos de hardware para entender o impacto de suas alterações, e itere o desenvolvimento para maximizar os dispositivos compatíveis.

Recursos

Documentos relacionados:

- O que é o AWS Device Farm?

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 práticas e ofertas 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 garantias, declarações ou condições de nenhum tipo, explícitas ou implícitas. 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.