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: 3 de outubro de 2023 (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 terá acesso a práticas recomendadas de arquitetura para projetar e operar sistemas confiáveis, seguros, eficientes, econômicos e sustentáveis 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 práticas recomendadas 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 sucesso 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 práticas recomendadas e principais estratégias para a arquitetura de sistemas na nuvem.

O AWS Well-Architected Framework documenta um conjunto de perguntas fundamentais que ajudam a compreender se uma arquitetura específica se alinha bem às práticas recomendadas da nuvem. Ele 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, além de fornecer 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. O Ferramenta AWS Well-Architected (Ferramenta AWS WA) é um serviço na nuvem que fornece um processo consistente para analisar e medir a arquitetura usando o AWS Well-Architected Framework. A AWS WA Tool 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

Name (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.

Name (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 dos custos	A capacidade de executar sistemas para entregar o valor empresarial ao menor preço.
Sustentabilidade	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, além da 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 workloads, 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 essenciais à missão, 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 on-premises, os clientes geralmente têm uma equipe central de arquitetura de tecnologia que atua como uma sobreposição para outras equipes de produtos ou atributos para verificar se estão seguindo as práticas recomendadas. As equipes de arquitetura de tecnologia tipicamente 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 entre equipes, em vez de termos uma equipe centralizada com esses recursos. Existem riscos na escolha de distribuir autoridade para tomada de decisões, por exemplo, verificar se as equipes atendem aos padrões internos. Atenuamos esses riscos de duas formas. Primeiro, nós temos práticas (processos, padrões, normas aceitas e formas de fazer as coisas) que se concentram em permitir que cada equipe tenha essa capacidade, e utilizamos especialistas que verificam se as equipes elevam o nível dos padrões que elas precisam cumprir. Segundo, implementamos mecanismos que realizam verificações automatizadas para verificar se os padrões estão sendo 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 práticas recomendadas. Para ajudar as novas equipes a obter essas capacidades ou as equipes existentes a elevar seus padrões, ativamos 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 engenheiros-chefe trabalha para que as práticas recomendadas sejam 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 práticas recomendadas, mas também usamos especialistas, como engenheiros-chefes, para defini-las. À medida que os engenheiros-chefes veem surgir novas práticas recomendadas, eles trabalham como uma comunidade para verificar se elas estão sendo seguidas pelas equipes. Com o tempo, essas práticas recomendadas 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 workloads, permitirão uma melhor compreensão dos riscos no 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 on-premises.

- Automatize com a experimentação da arquitetura em mente: a automação permite criar e replicar as workloads por custos baixos e evitar as despesas do trabalho manual. Você pode acompanhar as alterações em sua automação, auditar o impacto e reverter para os parâmetros anteriores quando necessário.
- Pense em 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 workload. 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 práticas recomendadas e das perguntas. Você pode encontrar orientações prescritivas sobre implementação no whitepaper Pilar Excelência operacional.

Tópicos

- Princípios de design
- Definição
- Práticas recomendadas
- Recursos

Princípios de design

Os princípios de design para obter a excelência operacional na nuvem são:

- Executar operações como código: na nuvem, você pode aplicar a todo o ambiente a mesma disciplina de engenharia usada para o código da aplicação. É possível definir toda a sua workload (aplicações, infraestrutura etc.) como código e atualizá-la com código. Você pode criar um script dos seus procedimentos de operações e automatizar o processo acionando-o em resposta a eventos. Ao executar operações como código, você limita o erro humano e cria respostas consistentes aos eventos.
- Fazer mudanças frequentes, pequenas e reversíveis: projete workloads escaláveis e com acoplamento fraco para permitir que os componentes sejam atualizados regularmente. Técnicas de implantação automatizadas, bem como mudanças menores e incrementais, reduzem o raio de expansão e permitem uma reversão mais rápida se ocorrerem falhas. Isso aumenta a confiança na entrega de mudanças benéficas à workload, mantendo a qualidade e possibilitando uma rápida adaptação às mudanças nas condições do mercado.
- Refinar os procedimentos operacionais com frequência: à medida que as workloads evoluem, expanda suas operações de forma adequada. ao usar os procedimentos de operação, procure oportunidades para melhorá-los. Organize análises regularmente e valide se todos os procedimentos estão em vigor e se as equipes estão familiarizadas com eles. Ao identificar lacunas, atualize os procedimentos adequadamente. Comunique as atualizações dos procedimentos a todas as partes interessadas e equipes. Promova o aprendizado gamificado em suas operações para compartilhar as práticas recomendadas e instruir as equipes.
- Antecipar falhas: execute exercícios “pre-mortem” para identificar possíveis origens de falhas, para que elas possam ser removidas ou mitigadas. Teste cenários de falha e valide sua compreensão do impacto deles. Teste os procedimentos de resposta para garantir que sejam eficazes e que as equipes estejam familiarizadas com o processo. Organize dias de jogo periódicos para testar workloads da equipe a eventos simulados.
- Aprender com todas as falhas operacionais: promova melhorias por meio de lições aprendidas com todos os eventos e falhas operacionais. Compartilhe o que foi aprendido com as equipes e a organização inteira.
- Usar serviços gerenciados: reduza a carga operacional usando serviços gerenciados da AWS sempre que possível. Crie procedimentos operacionais em torno das interações com esses serviços.
- Implementar a observabilidade para obter insights práticos: obtenha uma compreensão abrangente do comportamento, do desempenho, da confiabilidade, do custo e da integridade

da workload. Estabeleça indicadores-chave de desempenho (KPIs) e aproveite a telemetria de observabilidade para tomar decisões fundamentadas e agir imediatamente quando os resultados obtidos estiverem em risco. Melhore proativamente o desempenho, a confiabilidade e o custo com base em dados de observabilidade úteis.

Definição

Existem quatro áreas de práticas recomendadas 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 workload deve emitir as informações necessárias para apoiá-la. A implementação de serviços para permitir a integração, a implantação e a entrega de sua workload criará um fluxo maior de alterações benéficas na produção por meio da automação de processos repetitivos.

Pode haver riscos inerentes à operação da workload. Compreenda esses riscos e tome uma decisão embasada para entrar na produção. Suas equipes devem ser capazes de oferecer suporte à sua workload. As métricas operacionais e de negócios derivadas dos resultados de negócios desejados permitirão que você compreenda a integridade da workload e das atividades operacionais enquanto você responde 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 workload.

Práticas recomendadas

Tópicos

- Evoluir

Organização

Suas equipes devem ter um entendimento comum de toda a sua workload, da função que desempenham nela 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 verificará se você tem 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, confirme se você aplicou a devida diligência 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 abordar o risco.

Suas equipes devem compreender o papel delas na obtenção de resultados empresariais. As equipes devem entender o papel delas no êxito de outras equipes e a função das outras equipes no êxito delas, além de ter objetivos comuns. 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 workloads na sua organização.

Certifique-se de que haja proprietários identificados para cada componente de aplicação, workload, 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. A liderança sênior deve ser a patrocinadora, a defensora e a motivadora da adoção das práticas recomendadas e da evolução da organização. Permita que os membros da equipe tomem 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 estruturado e dedicado para o aprendizado. Garanta que os membros da equipe tenham os recursos (tanto ferramentas quanto pessoas) para serem bem-sucedidos e escalar para auxiliar os resultados empresariais. Aproveite a diversidade entre organizações para buscar várias perspectivas únicas. Use essa abordagem para aumentar a inovação, desafiar suas suposições e reduzir o risco de viés de confirmação. Aumente a inclusão, a diversidade e a acessibilidade em suas equipes para obter perspectivas benéficas.

Se houver requisitos externos de regulamentação ou conformidade aplicáveis à sua organização, use os recursos fornecidos pela Conformidade com a nuvem AWS para ajudar a instruir suas equipes de modo que elas possam determinar o impacto em suas prioridades. O Well-Architected Framework enfatiza o aprendizado, a medição e a melhoria. Ele 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, otimização de custos e sustentabilidade 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 (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 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.

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 ajudam a atender os 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 práticas recomendadas de excelência operacional, consulte o Appendix.)

OPS 1: How do you determine what your priorities are?

Everyone must understand their part in achieving business success. Have shared goals in order to set priorities for resources. This will maximize the benefits of your efforts.

OPS 2: How do you structure your organization to support your business outcomes?

Your teams must understand their part in achieving business outcomes. Teams must understand their roles in the success of other teams, the role of other teams in their success, and have shared goals. Understanding responsibility, ownership, how decisions are made, and who has authority to make decisions will help focus efforts and maximize the benefits from your teams.

OPS 3: How does your organizational culture support your business outcomes?

Provide support for your team members so that they can be more effective in taking action and supporting your business outcome.

Em determinado momento, talvez você queira 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 sucesso 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 workloads 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 workload 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 observabilidade e à investigação de problemas. A observabilidade vai além do simples monitoramento, fornecendo uma compreensão abrangente do funcionamento interno de um sistema com base em suas saídas externas. Baseada em métricas, logs e rastreamentos, a observabilidade oferece insights profundos sobre o comportamento e a dinâmica do sistema. Com uma observabilidade eficaz, as equipes podem discernir padrões, anomalias e tendências, permitindo que abordem proativamente possíveis problemas e mantenham a integridade ideal do sistema. Identificar os indicadores-chave de performance (KPIs) é fundamental para garantir o alinhamento entre as atividades de monitoramento e os objetivos de negócios. Esse alinhamento garante que as equipes tomem decisões baseadas em dados usando métricas que realmente importam, otimizando o desempenho do sistema e os resultados comerciais. Além disso, a observabilidade capacita as empresas a serem proativas em vez de reativas. As equipes podem entender as relações de causa e efeito em seus sistemas, prevendo e prevenindo problemas em vez de apenas reagir a eles. À medida que as workloads evoluem, é essencial revisitar e refinar a estratégia de observabilidade, garantindo que ela permaneça relevante e eficaz.

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 e correção mais rápidas, 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 workload, processos, procedimentos e pessoal para compreender os riscos operacionais relacionados à workload. Use um processo consistente (incluindo listas de verificação manuais ou automatizadas) para saber quando você estiver pronto para trabalhar com sua workload ou fazer uma mudança. Isso também ajudará a encontrar as áreas que você deve 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 e 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 da aplicação 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 workload, 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: How do you implement observability in your workload?

Implement observability in your workload so that you can understand its state and make data-driv en decisions based on business requirements.

OPS 5: How do you reduce defects, ease remediation, and improve flow into production?

Adopt approaches that improve flow of changes into production that achieve refactoring fast feedback on quality, and bug fixing. These accelerate beneficial changes entering production, limit issues deployed, and achieve rapid identification and remediation of issues introduced through deployment activities.

OPS 6: How do you mitigate deployment risks?

Adopt approaches that provide fast feedback on quality and achieve rapid recovery from changes that do not have desired outcomes. Using these practices mitigates the impact of issues introduce d through the deployment of changes.

OPS 7: How do you know that you are ready to support a workload?

Evaluate the operational readiness of your workload, processes and procedures, and personnel to understand the operational risks related to your workload.

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 identificar seus recursos. Identifique seus recursos de organização, contabilidade de custos e controles de acesso pensando na 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 workloads, planeje o que você fará com sistemas ativos que não estejam mais em conformidade.

Operar

A observabilidade permite que você se concentre em dados significativos e entenda as interações e os resultados da sua workload. Ao se concentrar em informações essenciais e eliminar dados desnecessários, você mantém uma abordagem direta para entender o desempenho da workload. É essencial não apenas coletar dados, mas também interpretá-los corretamente. Defina linhas de base claras e limites de alerta apropriados e monitore ativamente quaisquer desvios. Uma mudança em uma métrica-chave, especialmente quando correlacionada com outros dados, pode identificar áreas problemáticas específicas. Com a observabilidade, você está mais bem equipado para prever e enfrentar possíveis desafios, garantindo que sua workload opere sem problemas e atenda às necessidades de negócios.

A operação bem-sucedida de uma workload 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 processos 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 workloads 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 para identificar 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: How do you utilize workload observability in your organization?

Ensure optimal workload health by leveraging observability. Utilize relevant metrics, logs, and traces to gain a comprehensive view of your workload's performance and address issues efficient ly.

OPS 9: How do you understand the health of your operations?

Define, capture, and analyze operations metrics to gain visibility to operations events so that you can take appropriate action.

OPS 10: How do you manage workload and operations events?

Prepare and validate procedures for responding to events to minimize their disruption to your workload.

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

Aprenda, compartilhe e melhore continuamente para manter a excelência operacional. Dedique ciclos de trabalho para fazer melhorias incrementais quase 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 workload e os procedimentos operacionais.

Inclua ciclos de feedback 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: How do you evolve operations?

Dedicate time and resources for nearly continuous incremental improvement to evolve the effectiveness and efficiency of your operations.

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

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
- Gerenciamento de identidade e acesso
- Detecção
- Proteção de infraestrutura
- Proteção de dados
- Resposta a incidentes

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 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

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

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 apresenta uma visão geral dos princípios de design, das práticas recomendadas e das perguntas. Você pode encontrar orientações prescritivas sobre implementação no whitepaper Pilar 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:

- Democratize tecnologias avançadas: facilite a implementação de tecnologia avançada para a sua equipe delegando tarefas complexas ao seu fornecedor de nuvem. Em vez de solicitar que sua equipe de TI aprenda 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 ela se concentre no desenvolvimento de produtos, em vez de provisionamento e gerenciamento de recursos.
- Tenha alcance global em poucos minutos: a implantação de sua workload em várias regiões da AWS em todo o mundo permite oferecer menor latência e uma melhor experiência para seus clientes a um custo mínimo.
- Use arquiteturas sem servidor: as arquiteturas sem servidor eliminam a necessidade de executar e manter servidores físicos para realizar atividades tradicionais de computação. Os serviços de armazenamento sem servidor, por exemplo, podem atuar como sites estáticos (eliminando a necessidade de servidores 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.
- Experimente com mais frequência: com recursos virtuais e automatizáveis, você pode executar rapidamente testes comparativos usando diferentes tipos de instâncias, armazenamento ou configurações.
- Considere a solidariedade mecânica: entenda como os serviços de nuvem são consumidos e use sempre a abordagem tecnológica alinhada às suas metas de workload. Por exemplo, avalie padrões de acesso a dados ao selecionar abordagens de banco de dados ou armazenamento.

Definição

Existem cinco áreas de práticas recomendadas para eficiência de performance na nuvem:

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

Adote uma abordagem 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.

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

Práticas recomendadas

Tópicos

Seleção de arquitetura

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

Os recursos da AWS estão disponíveis em vários tipos e configurações, facilitando encontrar uma abordagem que atenda melhor às suas necessidades. Você também pode encontrar opções que não são facilmente obtidas com infraestrutura on-premises. 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 práticas recomendadas sobre eficiência de performance, consulte o Appendix.).

PERF 1: How do you select appropriate cloud resources and architecture patterns for your workload?

Often, multiple approaches are required for more effective performance across a workload. Well-Architected systems use multiple solutions and features to improve performance.

Computação e hardware

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

Na AWS, a computação é disponibilizada em três formatos: instâncias, contêineres e funções:

- Instâncias são servidores virtualizados cujos recursos podem ser alterados com um botão ou uma chamada de API. Como as decisões de recursos na nuvem não são imutáveis, você pode testar diferentes tipos de servidores. Na AWS, essas instâncias de servidor virtual vêm em diferentes famílias e tamanhos e oferecem uma ampla variedade de capacidades, inclusive unidades de estado sólido (SSDs) e unidades de processamento gráfico (GPUs).
- Contêineres são um método de virtualização do sistema operacional que permite executar uma aplicação e suas dependências em processos isolados por recursos. O AWS Fargate é um serviço de computação sem servidor para contêineres, mas 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 aplicar. Por exemplo, 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: How do you select and use compute resources in your workload?

The more efficient compute solution for a workload varies based on application design, usage patterns, and configuration settings. Architectures can use different compute solutions for various components and turn on different features to improve performance. Selecting the wrong compute solution for an architecture can lead to lower performance efficiency.

Gerenciamento de dados

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

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 de tecnologia 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 aplicações de empresas em todo o mundo.
- 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 workloads que exigem armazenamento persistente acessível por instâncias do EC2 e que ajuda você a ajustar aplicações com os níveis ideais de capacidade de armazenamento, performance e custo.
- 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 (Amazon 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 eficiente e econômico o processo de execução de sistemas de arquivos conhecidos,

para que você possa aproveitar os conjuntos de atributos 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: How do you store, manage, and access data in your workload?

The more efficient storage solution for a system varies based on the kind of access operation (block, file, or object), patterns of access (random or sequential), required throughput, frequency of access (online, offline, archival), frequency of update (WORM, dynamic), and availability and durability constraints. Well-architected systems use multiple storage solutions and turn on different features to improve performance and use resources efficiently.

Rede e entrega de conteúdo

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

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

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

PERF 4: How do you select and configure networking resources in your workload?

This question includes guidance and best practices to design, configure, and operate efficient networking and content delivery solutions in the cloud.

Processo e cultura

Ao arquitetar workloads, há princípios e práticas que você pode adotar para ajudar na melhor execução de workloads de nuvem eficientes e de alto desempenho. Para adotar uma cultura que promova a eficiência do desempenho das workloads na nuvem, considere estes princípios e práticas fundamentais:

Considere estes princípios fundamentais para construir essa cultura:

- Infraestrutura como código: defina sua infraestrutura como código usando abordagens como modelos do AWS CloudFormation. O uso de modelos permite que você coloque a infraestrutura no controle de origem junto com o código e as configurações de sua aplicação. Isso permite aplicar à sua infraestrutura as mesmas práticas usadas para desenvolver software, possibilitando uma iteração rápida.
- Pipeline de implantação: use um pipeline de Integração/Implantação Contínuas (CI/CD) (p. ex., repositório de código-fonte, sistemas de compilação, implantação e automação de teste) para implantar sua infraestrutura. Isso permite que você implante de maneira repetível, consistente e econômica enquanto itera.
- Métricas bem-definidas: configure e monitore métricas para capturar os indicadores-chave de performance (KPIs). Recomendamos o uso tanto de métricas técnicas quanto de negócios. Para aplicativos móveis ou sites, as principais métricas são a captura do tempo até o primeiro byte ou renderização. Outras métricas geralmente aplicáveis incluem contagem de thread, taxa de coleta de resíduos e estados de espera. Métricas de negócio, como o custo cumulativo agregado por solicitação, podem alertá-lo sobre maneiras de reduzir os custos. Considere com cuidado como você planeja interpretar as métricas. Por exemplo, você poderia escolher o máximo ou o 99º percentil, em vez da média.
- Teste a performance automaticamente: como parte de seu processo de implantação, acione automaticamente testes de performance após a aprovação bem-sucedida dos testes de execução mais rápidos. A automação deve criar um novo ambiente, configurar as condições iniciais, como dados de teste, e então executar uma série de benchmarks e testes de carga. Os resultados desses testes então devem ser vinculados de volta à compilação para que você possa acompanhar as mudanças de performance ao longo do tempo. Para testes de execução longa, você pode tornar essa parte do pipeline assíncrona do restante da compilação. Como alternativa, você pode realizar testes de performance durante a noite usando instâncias spot do Amazon EC2.
- Geração de carga: você deve criar uma série de scripts de teste que repliquem jornadas sintéticas ou pré-gravadas do usuário. Esses scripts devem ser idempotentes e não acoplados, e talvez você precise incluir scripts de pré-aquecimento para gerar resultados válidos. Seus scripts de teste

devem replicar tanto quanto for possível o comportamento do uso na produção. É possível usar soluções de software ou Software como Serviço (SaaS) para gerar a carga. Considere o uso das soluções do AWS Marketplace e de Instâncias spot: elas podem representar maneiras econômicas de gerar a carga.

- Visibilidade de performance: as métricas principais devem estar visíveis à sua equipe, especialmente métricas relacionadas a cada versão de compilação. Isso permite que você veja qualquer tendência positiva ou negativa importante ao longo do tempo. Você também deve exibir métricas do número de erros ou exceções para garantir que esteja testando um sistema em funcionamento.
- Visualização: use técnicas de visualização que deixem claro onde os problemas de performance, hot spots, estados de espera ou baixa utilização estão ocorrendo. Sobreponha métricas de performance a diagramas de arquitetura. Código ou gráficos de chamada podem ajudar a identificar problemas rapidamente.
- Processo de análise regular: arquiteturas com baixa performance geralmente são o resultado de um processo de análise de performance inexistente ou problemático. Se sua arquitetura está funcionando mal, a implementação de um processo de análise de desempenho permite que você promova melhorias iterativas.
- Otimização contínua: adote uma cultura para otimizar continuamente a eficiência da performance da workload na nuvem.

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

PERF 5: What process do you use to support more performance efficiency for your workload?

When architecting workloads, there are principles and practices that you can adopt to help you better run efficient high-performing cloud workloads. To adopt a culture that fosters performance efficiency of cloud workloads, consider these key principles and practices.

Recursos

Consulte os seguintes recursos para saber mais sobre nossas melhores práticas para eficiência de performance.

Documentação

- Amazon S3 Otimização da performance

- Amazon EBS Performance de volume

Whitepaper

- Pilar Eficiência de performance

Vídeo

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

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
- Definição
- Práticas recomendadas
- Recursos

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 administrativa e operacional, liberando você para trabalhar em aplicativos e atividades relacionadas a negócios.

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

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

COST 7: Como você usa 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

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

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 principal	Reestruturação importante do pilar de performance para elevar o número de áreas de práticas recomendadas para cinco. Grande atualizaç ão das práticas recomenda das e orientações no pilar de segurança em Resposta a incidentes (SEC 10). Principai s mudanças de conteúdo e consolidação em áreas de excelência operacional OPS 04, 05, 06, 08 e 09. Atualizaç ões de orientação em todos os pilares de otimização de custos e confiabilidade . Atualizações secundárias nos níveis de risco do pilar de sustentabilidade.	October 3, 2023
Atualizações para o novo Framework	Práticas recomendadas atualizadas com orientações prescritivas e novas práticas recomendadas adicionadas. Novas perguntas adicionadas aos pilares de Segurança e Otimização de Custos.	April 10, 2023
Atualização secundária	Adição da definição de nível de esforço e atualização das	October 20, 2022

	práticas recomendadas no apêndice.	
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	Remoção 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 o AWS Well-Architected Labs e AWS Well-Architected Partners, além de correções secundárias para possibili tar uma versão em várias linguagens do Framework.	July 1, 2019

Whitepaper atualizado	Revisão e reescrita da maioria das perguntas e respostas, para garantir que as perguntas se concentrem em um tópico de cada vez. Isso fez com que algumas perguntas anteriores fossem divididas em várias perguntas. Adição de termos comuns às definições (carga de trabalho, componente etc). Apresentação alterada da pergunta no corpo principal para incluir texto descritivo.	November 1, 2018
Whitepaper atualizado	Atualizações para simplificar o texto de pergunta, padronizar 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 atualizados para refletir a evolução da AWS.	November 1, 2017
Whitepaper atualizado	Atualização do Framework para incluir o pilar de excelência operacional e revisão e atualização dos outros pilares para reduzir a duplicação e incorporar aprendizados da realização de análises com milhares de clientes.	November 1, 2016

Atualizações secundárias	Atualização do Apêndice com informações atuais do Amazon CloudWatch Logs.	November 1, 2015
Publicação inicial	Publicação do AWS Well-Arch itected Framework.	October 1, 2015

Apêndice: Perguntas e práticas recomendadas

Este apêndice resume todas as perguntas e práticas recomendadas do AWS Well-Architected Framework.

Pilares

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

Excelência operacional

O pilar Excelência operacional inclui a capacidade de oferecer conformidade com o desenvolvimento e de executar workloads com eficácia, obter insights sobre as operações e melhorar continuamente processos e procedimentos de suporte para oferecer valor empresarial. Você pode encontrar orientações prescritivas sobre implementação no whitepaper sobre o pilar de excelência operacional.

Áreas de práticas recomendadas

- Organização
- Preparar
- Operar
- Evoluir

Organização

Perguntas

- OPERAÇÕES 1. Como determinar quais são suas prioridades?
- OPERAÇÕES 2. Como estruturar sua organização para dar suporte aos resultados comerciais?
- OPERAÇÕES 3. Como sua cultura organizacional oferece suporte aos resultados comerciais?

OPERAÇÕES 1. Como determinar quais são suas prioridades?

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

Práticas recomendadas

- OPS01-BP01 Avaliar as necessidades dos clientes externos
- 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

Governança refere-se a um conjunto de políticas, regras ou frameworks que uma empresa usa para atingir metas de negócios. Os requisitos de governança são gerados dentro da organização. Eles podem afetar os tipos de tecnologia que você escolhe ou influenciar a maneira como você opera sua workload. Incorpore requisitos de governança organizacional em sua workload. Conformidade é a capacidade de demonstrar que você implementou os requisitos de governança.

Resultado desejado:

- Os requisitos de governança são incorporados ao design arquitetural e à operação da workload.
- Você pode fornecer prova de que seguiu os requisitos de governança.
- Os requisitos de governança são revistos e atualizados regularmente.

Antipadrões comuns:

- Sua organização exige que a conta raiz tenha autenticação multifator. Você não implementa esse requisito e a conta raiz é comprometida.
- Durante o design da workload, você escolhe um tipo de instância que não é aprovada pelo departamento de TI. Você não consegue iniciar a workload e precisa começar a reprojetá-la.
- É obrigatório que você tenha um plano de recuperação de desastres. Você não cria um, e a workload sofre uma interrupção prolongada.
- Sua equipe quer usar novas instâncias, mas seus requisitos de governança não foram atualizados para permiti-las.

Benefícios do estabelecimento desta prática recomendada:

- A aderência aos requisitos de governança alinha sua workload às políticas da organização como um todo.
- Os requisitos de governança refletem os padrões e as práticas recomendas do setor para sua organização.

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

Orientações para a implementação

Identifique o requisito de governança trabalhando com as partes interessadas e as organizações de governança. Inclua os requisitos de governança em sua workload. Prepare-se para demonstrar prova de que você seguiu os requisitos de governança.

Exemplo de clientes

Na Loja UmaEmpresa, a equipe de operações em nuvem trabalha com as partes interessadas dentro da organização para desenvolver requisitos de governança. Por exemplo, eles proíbem acesso SSH a instâncias do Amazon EC2. Caso as equipes precisem de acesso ao sistema, elas devem usar o AWS Systems Manager Session Manager. A equipe de operações em nuvem atualiza regularmente os requisitos de governança à medida que novos serviços são disponibilizados.

Etapas da implementação

1. Identifique as partes interessadas referentes à sua workload, incluindo quaisquer equipes centralizadas.
2. Trabalhe com as partes interessadas para identificar requisitos de governança.
3. Assim que gerar uma lista, priorize os itens de melhoria e comece a implementá-los na workload.
   a. Use serviços como o AWS Config para criar governança como código e validar se esses requisitos de governança são seguidos.
   b. Se usar o AWS Organizations, poderá utilizar políticas de controle de serviços para implementar requisitos de governança.
4. Forneça uma documentação que valide a implementação.

Nível de esforço do plano de implementação: médio. A implementação de requisitos de governança pode exigir a reformulação de sua workload.

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP04 Avaliar os requisitos de conformidade: a conformidade é como a governança, mas provém de fora da organização.

Documentos relacionados:

- Guia de gerenciamento e governança do ambiente de nuvem da AWS

- Práticas recomendadas para políticas de controle de serviço do AWS Organizations
- Governança na Nuvem AWS: o equilíbrio entre agilidade e segurança
- O que é governança, risco e conformidade (GRC)?

Vídeos relacionados:

- Gerenciamento e governança da AWS: configuração, conformidade e auditoria – AWS Online Tech Talks
- AWS re:Inforce 2019: Governança da era da nuvem (DEM12-R1)
- AWS re:Invent 2020: Como alcançar a conformidade usando o AWS Config
- AWS re:Invent 2020: Governança ágil na AWS GovCloud (US)

Exemplos relacionados:

- Amostras de pacote de conformidade do AWS Config

Serviços relacionados:

- AWS Config
- AWS Organizations: políticas de controle de serviços

OPS01-BP04 Avaliar os requisitos de conformidade

Os requisitos de conformidade normativos, setoriais e internos são um importante motivador para definir as prioridades de sua organização. Seu framework de conformidade pode impedir você de usar tecnologias ou localizações geográficas específicas. Realize a devida diligência se não for identificado nenhum framework de conformidade externo. Gere auditorias ou relatórios que validem a conformidade.

Se você anunciar que seu produto atende a padrões de conformidade específicos, deverá ter um processo interno para garantir a conformidade contínua. Os exemplos de padrões de conformidade incluem o PCI DSS, o FedRAMP e a HIPAA. Os padrões de conformidade aplicáveis são determinados por vários fatores, por exemplo, quais tipos de dados a solução armazena ou transmite e a quais regiões a solução oferece suporte.

Resultado desejado:

- Os requisitos de conformidade normativos, setoriais e internos são incorporados na seleção arquitetural.
- Você pode validar a conformidade e gerar relatórios de auditoria.

Antipadrões comuns:

- Partes da workload enquadram-se no framework Padrão de Segurança de Dados do Setor de Cartões de Pagamento (PCI-DSS), mas a workload armazena dados de cartões de crédito não criptografados.
- Seus desenvolvedores e arquitetos de software não estão a par do framework de conformidade que sua organização deve adotar.
- A auditoria anual de Controle de Sistemas e Organizações: Tipo II (SOC2) será feita em breve e você não consegue verificar se esses controles estão em vigor.

Benefícios do estabelecimento desta prática recomendada:

- Avaliar e compreender os requisitos de conformidade que se aplicam à sua workload informará como você prioriza seus esforços para entregar valor empresarial.
- Você escolhe as localizações e tecnologias corretas, que são congruentes com seu framework de conformidade.
- Quando a workload é projetada para ser auditável, você tem a possibilidade de provar que está seguindo seu framework de conformidade.

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

Orientações para a implementação

Implementar essa prática recomendada significa incorporar os requisitos de conformidade no processo de design da arquitetura. Os membros de sua equipe estão a par do framework de conformidade necessário. Você valida a conformidade de acordo com o framework.

Exemplo de clientes

A Loja UmaEmpresa armazena informações de cartão de crédito dos clientes. Os desenvolvedores da equipe de armazenamento de cartões sabem que eles precisam acatar o framework PCI-DSS. Eles tomaram medidas para verificar que as informações de cartão de crédito são armazenadas e

acessadas com segurança, de acordo com o framework PCI-DSS. Todo ano, eles trabalham com a equipe de segurança para validar a conformidade.

Etapas da implementação

1. Trabalhe com as equipes de segurança e governança para determinar quais frameworks de conformidade normativos, setoriais ou internos a workload deve seguir. Incorpore os frameworks de conformidade em sua workload.
   a. Valide a conformidade contínua dos recursos da AWS com serviços como o AWS Compute Optimizer e o AWS Security Hub.
2. Instrua os membros da equipe sobre os requisitos de conformidade para que possam operar e expandir a workload de acordo com eles. Os requisitos de conformidade devem ser incluídos nas escolhas de arquitetura e tecnologia.
3. Dependendo do framework de conformidade, pode ser necessário gerar um relatório de auditoria ou conformidade. Trabalhe com sua organização para automatizar esse processo o máximo possível.
   a. Use serviços como o AWS Audit Manager para validar a conformidade e gerar relatórios de auditoria.
   b. Você pode baixar documentos de segurança e conformidade da AWS com o AWS Artifact.

Nível de esforço do plano de implementação: médio. A implementação de frameworks de conformidade pode ser um desafio. A geração de relatórios de auditoria e de documentos de conformidade aumenta ainda mais complexidade.

Recursos

Práticas recomendadas relacionadas:

- SEC01-BP03 Identificar e validar objetivos de controle: os objetivos de controle de segurança são uma parte importante da conformidade geral.
- SEC01-BP06 Automatizar testes e validar controles de segurança em pipelines: como parte de seus pipelines, validade os controles de segurança. Você também pode gerar documentação de conformidade para novas mudanças.
- SEC07-BP02 Definir controles de proteção de dados: muitos frameworks de conformidade têm políticas baseadas em processamento e armazenamento de dados.
- SEC10-BP03 Preparar recursos forenses: às vezes é possível usar recursos forenses em auditoria de conformidade.

Documentos relacionados:

- Centro de Conformidade da AWS
- Recursos de conformidade da AWS
- Whitepaper AWS: risco e conformidade
- Modelo de Responsabilidade Compartilhada da AWS
- Serviços da AWS no escopo por programa de conformidade

Vídeos relacionados:

- AWS re:Invent 2020:Como alcançar a conformidade usando o AWS Compute Optimizer
- AWS re:Invent 2021: Conformidade, garantia e auditoria
- AWS Summit ATL 2022: Implementação de conformidade, garantia e auditoria na AWS (COP202)

Exemplos relacionados:

- PCI DSS e Práticas recomendadas de segurança básica da AWS na AWS

Serviços relacionados:

- AWS Artifact
- AWS Audit Manager
- AWS Compute Optimizer
- AWS Security Hub

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.

OPERAÇÕES 2. Como estruturar sua organização para dar suporte aos resultados comerciais?

Suas equipes devem compreender o papel delas na obtenção de resultados empresariais. As equipes devem entender o papel delas no êxito de outras equipes, 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

Os recursos para sua workload devem ter proprietários identificados para controle de alterações, resolução de problemas e outras funções. Atribuem-se proprietários para workloads, contas, infraestrutura, plataformas e aplicações. A propriedade é registrada usando ferramentas como um registro central ou metadados anexados aos recursos. O valor empresarial dos componentes indica os processos e procedimentos aplicados a eles.

Resultado desejado:

- Os recursos têm proprietários identificados usando metadados ou um registro central.
- Os membros da equipe podem identificar quem é proprietários dos recursos.
- As contas têm um único proprietário quando possível.

Antipadrões comuns:

- Os contatos alternativos para suas Contas da AWS não estão preenchidos.
- Os recursos não têm as tags que identificam as equipes às quais eles pertencem.
- Você tem uma fila ITSM sem mapeamento de e-mail.
- Duas equipes são proprietárias de uma mesma parte essencial da infraestrutura.

Benefícios do estabelecimento desta prática recomendada:

- O controle de alterações para recursos é fácil com a atribuição de propriedade.
- Você pode envolver os proprietários corretos na resolução de problemas.

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

Orientações para a implementação

Defina o que significa propriedade para os casos de uso de recursos em seu ambiente. A propriedade pode se referir a quem supervisiona alterações no recurso e apoia o recurso durante a resolução de problemas ou a quem é responsável pela parte financeira. Especifique e registre proprietários para recursos, incluindo nome, informações de contato, organização e equipe.

Exemplo de clientes

A Loja UmaEmpresa define propriedade como a equipe ou a pessoa proprietária das alterações e do suporte para os recursos. Eles utilizam o AWS Organizations para gerenciar as Contas da AWS. Os contatos de conta alternativos são configurados usando as caixas de entrada de grupo. Cada fila ITSM é mapeada para um alias de e-mail. As tags identificam quem é proprietário dos recursos da AWS. Para outras plataformas e infraestrutura, eles têm uma página de wiki que identifica informações sobre propriedade e contato.

Etapas da implementação

1. Para começar, identifique a propriedade sobre sua organização. A propriedade pode estar relacionada a quem é proprietário do risco referente ao recurso, a quem é proprietário das alterações referentes ao recurso ou a quem apoia o recurso na resolução de problemas. Propriedade também pode significar propriedade financeira ou administrativa pelo recurso.
2. Use o AWS Organizations para gerenciar contas. Você pode gerenciar contatos alternativos centralmente para as suas contas.
    a. O uso de endereços de e-mail ou de números de telefones de propriedade da empresa para informações de contato permite acessá-los mesmo quando os indivíduos aos quais eles pertencem não estiverem mais na organização. Por exemplo, crie listas de distribuição de e-mail separadas para faturamento, operações e segurança, e configure-as como contatos de Faturamento, Segurança e Operações em cada Conta da AWS ativa. Várias pessoas receberão notificações da AWS e poderão respondê-las, mesmo que alguém esteja de férias, mude de função ou saia da empresa.
    b. Se uma conta não for gerenciada pelo AWS Organizations, os contatos alternativos para contas ajudarão a AWS a entrar em contato com o pessoal apropriado, se necessário. Configure os contatos alternativos da conta para apontar para um grupo em vez de uma pessoa.
3. Use tags para identificar proprietários de recursos da AWS. Você pode especificar os proprietários e as respectivas informações de contato em tags separadas.
    a. Pode usar regras do AWS Config para reforçar que os recursos têm as tags de propriedade necessárias.

   b. Para obter orientações detalhadas sobre como elaborar uma estratégia de marcação para sua organização, consulte o whitepaper Práticas recomendadas de marcação da AWS.
4. Para outros recursos, plataformas e infraestrutura, crie uma documentação que identifique a propriedade. Ela deve ser acessível a todos os membros da equipe.

Nível de esforço do plano de implementação: baixo. Utilize informações de contato da conta e tags para atribuir propriedade a recursos da AWS. Para outros recursos, você pode usar algo simples como uma tabela em uma wiki para registrar a propriedade e informações de contato ou usar uma ferramenta de ITSM para mapear a propriedade.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP02 Processos e procedimentos com proprietários identificados: os processos e procedimentos para apoiar os recursos dependem do proprietário do recurso.
- OPS02-BP04 Os membros da equipe sabem pelo que são responsáveis: os membros da equipe devem saber de quais recursos eles são proprietários.
- OPS02-BP05 Existem mecanismos para identificar a responsabilidade e a propriedade: é necessário usar mecanismos como tags e contatos de conta para que a propriedade possa ser descoberta.

Documentos relacionados:

- AWS Account Management: Atualizar informações de contato
- Regras do AWS Config: required-tags
- AWS Organizations: Atualizar contatos alternativos em sua organização
- Whitepaper Práticas recomendadas de marcação AWS

Exemplos relacionados:

- Regras do AWS Config: Amazon EC2 com regras requeridas e valores válidos

Serviços relacionados:

- AWS Config

- AWS Organizations

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 fundamenta 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. Solicitações incluem adições, alterações e exceções. Essas solicitações passam por um processo de gerenciamento de alterações. Tome decisões embasadas para aprovar solicitações quando elas forem viáveis e foram consideradas apropriadas após uma avaliação de benefícios e riscos.

Resultado desejado:

- Você pode fazer solicitações para alterar processos, procedimentos e recursos com base na propriedade atribuída.
- As alterações são feitas de maneira deliberada, ponderando benefícios e riscos.

Antipadrões comuns:

- Você precisa atualizar a maneira como implanta sua aplicação, mas não há como solicitar uma alteração no processo de implantação à equipe de operações.
- O plano de recuperação de desastres deve ser atualizado, mas não há nenhum proprietário identificado para solicitar alterações no plano.

Benefícios do estabelecimento desta prática recomendada:

- Processos, procedimentos e recursos podem evoluir à medida que os requisitos mudam.
- Os proprietários podem tomar decisões embasadas sobre quando realizar alterações.
- As alterações são feitas de maneira deliberada.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientações para a implementação

Para implementar esta prática recomendada, você precisa estar em uma posição em que possa solicitar alterações em processos, procedimentos e recursos. O processo de gerenciamento de alterações pode ser simples. Documente o processo de gerenciamento de alterações.

Exemplo de clientes

A Loja UmaEmpresa usa a matriz de atribuição de responsabilidades (RACI) para identificar quem é proprietário das alterações em processos, procedimentos e recursos. Eles documentaram o processo de gerenciamento de alterações, que é simples e fácil de seguir. Usando a matriz RACI e o processo, qualquer pessoa pode enviar solicitações de alteração.

Etapas da implementação

1. Identifique processos, procedimentos e recursos para sua workload e os proprietários de cada um. Documente-os no sistema de gerenciamento de conhecimentos.
   a. Se você não tiver implementado OPS02-BP01 Recursos com proprietários identificados, OPS02-BP02 Processos e procedimentos com proprietários identificados ou OPS02-BP03

Atividades de operações com proprietários identificados responsáveis pela performance, comece com estes.

2. Trabalhe com as partes interessadas em sua organização para desenvolver um processo de gerenciamento de alterações. O processo deve abranger adições, alterações e exceções para recursos, processos e procedimentos.
   a. Você pode usar o Gerente de Alterações do AWS Systems Manager como plataforma de gerenciamento de alterações para recursos de workload.
3. Documente o processo de gerenciamento de alterações em seu sistema de gerenciamento de conhecimentos.

Nível de esforço do plano de implementação: médio. O desenvolvimento de um processo de gerenciamento de alterações deve estar alinhado com várias partes interessadas em sua organização.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP01 Recursos com proprietários identificados: para criar um processo de gerenciamento de alterações, primeiro é necessário identificar os proprietários dos recursos.
- OPS02-BP02 Processos e procedimentos com proprietários identificados: para criar um processo de gerenciamento de alterações, primeiro é necessário identificar os proprietários dos recursos.
- OPS02-BP03 Atividades de operações com proprietários identificados responsáveis pela performance: para criar um processo de gerenciamento de alterações, primeiro é necessário identificar os proprietários.

Documentos relacionados:

- AWS Prescriptive Guidance - Foundation playbook for AWS large migrations: Creating RACI matrices
- Whitepaper Gerenciamento de alterações na nuvem

Serviços relacionados:

- Gerente de Alterações do AWS Systems Manager

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). Os canais de comunicação entre equipes são documentados. Ao entender o impacto do trabalho das equipes nos resultados de negócios e nos resultados de outras equipes e organizações, você conhece a priorização de tarefas delas e as ajuda a responder 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.

Resultado desejado:

- Os acordos de trabalho ou apoio entre equipes são combinados e documentados.
- As equipes que apoiam ou trabalham umas com as outras definiram canais de comunicação e expectativas de resposta.

Antipadrões comuns:

- Ocorre um problema na produção e duas equipes separadas começam iniciam a resolução de problemas de maneira independente. Esses esforços isolados estendem a interrupção.
- A equipe de operações necessita de assistência da equipe de desenvolvimento, mas nenhum tempo de resposta foi acordado. A solicitação está parada em uma lista de pendências.

Benefícios do estabelecimento desta prática recomendada:

- As equipes sabem interagir e apoiar uma à outra.
- As expectativas quanto à capacidade de resposta são claras.
- Os canais de comunicação estão nitidamente definidos.

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

Orientações para a implementação

A implementação desta prática recomendada significa que não há ambiguidade quanto à forma como as equipes trabalham uma com a outra. Os acordos formais sistematizam de que maneira as

equipes trabalham juntas ou apoiam uma à outra. Os canais de comunicação entre as equipes são documentados.

Exemplo de clientes

A equipe de SRE da Loja UmaEmpresa tem um Acordo de Serviço com a equipe de desenvolvimento. Sempre que a equipe de desenvolvimento faz uma solicitação no sistema de tíquetes, ela pode esperar uma resposta em 15 minutos. Se não houver nenhuma interrupção no local, a equipe de SRE toma a dianteira na investigação e conta com o apoio da equipe de desenvolvimento.

Etapas da implementação

1. Trabalhando com as partes interessadas na organização, desenvolva acordos entre as equipes com base nos processos e procedimentos.
   a. Se um processo ou procedimento for compartilhado entre as duas equipes, desenvolva um runbook sobre como as equipes trabalharão juntas.
   b. Se houver dependências entre as equipes, estabeleça um SLA de resposta às solicitações.
2. Documente as responsabilidades no sistema de gerenciamento de conhecimentos.

Nível de esforço do plano de implementação: médio. Se não houver nenhum entendimento entre as equipes, pode ser difícil chegar a um acordo com as partes interessadas na organização.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP02 Processos e procedimentos com proprietários identificados: para estabelecer acordos entre as equipes, primeiro é necessário identificar o proprietário do processo.
- OPS02-BP03 Atividades de operações com proprietários identificados responsáveis pela performance: para estabelecer acordos entre as equipes, primeiro é necessário identificar o proprietário das atividades de operações.

Documentos relacionados:

- AWS Executive Insights: Impulsionando inovação e velocidade com as equipes de duas pizzas da Amazon
- Introdução a DevOps na AWS: Equipes de duas pizzas)

OPERAÇÕES 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.

Resultado desejado:

- A comunicação fornece contexto, detalhes e expectativas de tempo.
- Os membros da equipe têm um entendimento claro sobre quando e como agir em resposta a comunicações.
- Utilize calendários de alterações para chamar a atenção para as alterações previstas.

Antipadrões comuns:

- Um alerta falso-positivo é acionado várias vezes por semana. Você silencia a notificação toda vez em que ela ocorre.
- Você recebe uma solicitação para alterar seu grupo de segurança, mas não é passada nenhuma expectativa sobre quando isso deve ocorrer.
- Você recebe notificações constantes por chat quando a escala dos sistemas é aumentada verticalmente, mas nenhuma ação é necessária. Você evita o canal por chat e perde uma notificação importante.

- Fizeram uma alteração na produção sem informar a equipe de operações. A alteração aciona um alerta e a equipe de plantão é ativada.

Benefícios do estabelecimento desta prática recomendada:

- Sua organização evita a fadiga de alertas.
- Os membros da equipe podem agir com o contexto necessário e as expectativas indispensáveis.
- As alterações podem ser feitas durante períodos apropriados e reduzir o risco.

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

Orientações para a implementação

Para implementar esta prática recomendada, você precisa trabalhar com as partes interessadas na organização para ajustar padrões de comunicação. Divulgue esses padrões para toda a organização. Identifique e remova alertas falso-positivos ou que fiquem sempre ativos. Utilize calendários de alterações para que os membros da equipe saibam quando é possível agir e quais atividades estão pendentes. Verifique se as comunicações geram ações claras com o contexto necessário.

Exemplo de clientes

A Loja UmaEmpresa usa o chat como principal meio de comunicação. Alertas e outras informações são divulgados em canais específicos. Quando alguém precisa agir, o resultado desejado é claramente expresso, e, em muitos casos, as pessoas recebem um runbook ou manual para uso nessas situações. O calendário de alterações é usado para programar alterações importantes nos sistemas de produção.

Etapas da implementação

1. Analise os alertas para identificar falso-positivos ou alertas que são acionados constantemente. Remova ou altere esses alertas para que sejam acionados quando há necessidade de intervenção humana. Se um alerta for acionado, forneça um runbook ou manual.
   a. Você pode usar Documentos do AWS Systems Manager para criar manuais e runbooks para alertas.
2. Mecanismos estão em vigor para fornecer notificações de riscos ou eventos planejados de maneira clara e prática com aviso prévio em tempo suficiente para permitir respostas apropriadas. Use listas de e-mails ou canais por chat para enviar notificações antes dos eventos planejados.

   a. É possível usar o AWS Chatbot para enviar alertas e responder a eventos dentro da plataforma de mensagens de suas organizações.

3. Forneça uma fonte de informações acessível em que eventos planejados possam ser descobertos. Forneça notificações de eventos planejados oriundos do mesmo sistema.
   a. O Calendário de Alterações do AWS Systems Manager pode ser usado para criar períodos em que as alterações podem ocorrer. Isso oferece aos membros da equipe um aviso prévio sobre quando eles podem fazer alterações com segurança.
4. Monitore notificações de vulnerabilidade e informações de patches para identificar vulnerabilidades nos riscos reais e potenciais associados aos componentes da workload. Forneça uma notificação aos membros da equipe para que eles possam agir.
   a. Você pode assinar os boletins de segurança da AWS para receber notificações sobre vulnerabilidades na AWS.

Recursos

Práticas recomendadas relacionadas:

- OPS07-BP03 Usar runbooks para realizar procedimentos: para que as comunicações tenham efeito prático, forneça um runbook quando o resultado for conhecido.
- OPS07-BP04 Usar manuais para investigar problemas: quando não se conhece o resultado, os manuais podem tornar as comunicações acionáveis.

Documentos relacionados:

- Boletins de segurança da AWS
- Open CVE

Exemplos relacionados:

- Laboratórios do Well-Architected: Gerenciamento de inventário e patches (Nível 100)

Serviços relacionados:

- AWS Chatbot
- Calendário de alterações do AWS Systems Manager

- Documentos do AWS Systems Manager

OPS03-BP05 Incentivo à experimentação

A experimentação é um catalisador para transformar novas ideias em produtos e recursos. Ela acelera o aprendizado e mantém os membros da equipe interessados e envolvidos. Os membros da equipe são incentivados a experimentar com frequência para promover a inovação. Mesmo quando um resultado não desejado ocorre, é importante saber o que não se deve fazer. Os membros da equipe não são punidos por experimentos bem-sucedidos com resultados indesejados.

Resultado desejado:

- Sua organização incentiva a experimentação para promover a inovação.
- Os experimentos são usados como oportunidade de aprendizado.

Antipadrões comuns:

- Você deseja executar um teste A/B, mas não há nenhum mecanismo para conduzir o experimento. Você implanta uma alteração de interface do usuário sem a possibilidade de testá-la. O resultado é uma experiência negativa para o cliente.
- Sua empresa tem apenas o ambiente de preparação e produção. Como não há área restrita para testes para experimentar novos recursos ou produtos, você precisa realizar experimentos no ambiente de produção.

Benefícios do estabelecimento desta prática recomendada:

- A experimentação promove a inovação.
- Você reagir mais depressa ao feedback dos usuários por meio da experimentação.
- Sua organização desenvolve uma cultura de aprendizado.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientações para a implementação

Os experimentos devem ser conduzidos de maneira segura. Utilize vários ambientes para experimentar, sem colocar em risco os recursos da produção. Use testes A/B e sinalizadores de

recursos para testar experimentos. Ofereça aos membros da equipe a possibilidade de conduzir experimentos em um ambiente de área restrita para testes.

Exemplo de clientes

A Loja UmaEmpresa estimula a experimentação. Os membros da equipe podem usar 20% da semana de trabalho para experimentar ou aprender novas tecnologias. Eles têm um ambiente de área restrita para testes no qual podem inovar. São usados testes A/B para novos recursos com o objetivo de validá-los com um feedback de usuário real.

Etapas da implementação

1. Trabalhe com a liderança em toda a sua organização para favorecer a experimentação. Os membros da equipe devem ser incentivados a conduzir experimentos de maneira segura.
2. Ofereça aos membros da equipe um ambiente em que eles possa experimentar com segurança. Eles devem ter acesso a um ambiente semelhante ao de produção.
   a. Você pode usar uma Conta da AWS separada para criar um ambiente de área restrita para testes para experimentação. O AWS Control Tower pode ser usado para provisionar essas contas.
3. Use sinalizadores de recursos e testes A/B para experimentar com segurança e coletar feedback dos usuários.
   a. O AWS AppConfig Feature Flags oferece a possibilidade de criar sinalizadores de recursos.
   b. O Amazon CloudWatch Evidently pode ser usado para realizar testes A/B em uma implantação limitada.
   c. Você pode usar versões do AWS Lambda para implantar uma nova versão de uma função para testes beta.

Nível de esforço do plano de implementação: alto. A viabilização de um ambiente para experimentação e de uma maneira segura para os membros da equipe conduzirem experimentos pode exigir um investimento significativo. Você também pode precisar modificar o código da aplicação para usar sinalizadores de recursos ou respaldar testes A/B.

Recursos

Práticas recomendadas relacionadas:

- OPS11-BP02 Executar análise pós-incidente: aprender com incidentes é um motivador fundamental para a inovação, bem como a experimentação.

- OPS11-BP03 Implementar loops de feedback: os ciclos de feedback são um componente importante da experimentação.

Documentos relacionados:

- Uma visão interna da cultura da Amazon: experimentação, erro e obsessão pelo cliente
- Práticas recomendadas para criar e gerenciar contas em área restrita para testes na AWS
- Crie uma cultura de experimentação viabilizada pela nuvem
- Viabilização da experimentação e inovação na nuvem na SulAmérica Seguros
- Experimente mais, erre menos
- Organização do ambiente da AWS usando várias contas: UO de área restrita para testes
- Como usar o AWS AppConfig Feature Flags

Vídeos relacionados:

- AWS On Air ft. Amazon CloudWatch Evidently | AWS Eventos
- AWS On Air San Fran Summit 2022 ft. Integração do AWS AppConfig Feature Flags com o Jira
- AWS re:Invent 2022: Implantação não é lançamento: controle seus lançamentos com sinalizadores de recursos (BOA305-R)
- Crie programaticamente uma Conta da AWS com o AWS Control Tower
- Configuração de um ambiente de várias contas da AWS que usa as práticas recomendadas para AWS Organizations

Exemplos relacionados:

- AWS Innovation Sandbox
- Personalização básica de ponta a ponta para comércio eletrônico

Serviços relacionados:

- Amazon CloudWatch Evidently
- AWS AppConfig
- AWS Control Tower

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

- OPERAÇÕES 4. Como implementar a observabilidade em sua workload?
- OPERAÇÕES 5. Como reduzir defeitos, facilitar a correção e melhorar o fluxo na produção?
- OPERAÇÕES 6. Como reduzir os riscos de implantação?
- OPERAÇÕES 7. Como saber se está pronto para oferecer suporte a uma workload?

OPERAÇÕES 4. Como implementar a observabilidade em sua workload?

Implemente a observabilidade na workload para que você possa entender seu estado e tomar decisões baseadas em dados com base nos requisitos de negócios.

Práticas recomendadas

- OPS04-BP01 Identificar os indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS04-BP03 Implementar a telemetria da experiência do usuário
- OPS04-BP04 Implementar a telemetria de dependências
- OPS04-BP05 Implementar rastreamento distribuído

OPS04-BP01 Identificar os indicadores-chave de performance

A implementação da observabilidade em sua workload começa com a compreensão de seu estado e a tomada de decisões baseadas em dados conforme os requisitos de negócios. Uma das formas mais eficazes de garantir o alinhamento entre as atividades de monitoramento e os objetivos de negócios é definir e monitorar os indicadores-chave de performance (KPIs).

Resultado desejado: Práticas de observabilidade eficientes que estão estreitamente alinhadas aos objetivos de negócios, garantindo que os esforços de monitoramento estejam sempre a serviço de resultados comerciais tangíveis.

Antipadrões comuns:

- KPIs indefinidos: trabalhar sem KPIs claros pode levar ao monitoramento excessivo ou insuficiente, perdendo sinais vitais.

- KPIs estáticos: não revisitar ou refinar os KPIs à medida que a workload ou os objetivos de negócios evoluem.
- Desalinhamento: foco em métricas técnicas que não se correlacionam diretamente com os resultados comerciais ou são mais difíceis de correlacionar com problemas do mundo real.

Benefícios de estabelecer esta prática recomendada:

- Facilidade de identificação de problemas: os KPIs de negócios geralmente mostram os problemas com mais clareza do que as métricas técnicas. Uma queda em um KPI comercial pode identificar um problema com mais eficiência do que analisar várias métricas técnicas.
- Alinhamento comercial: garante que as atividades de monitoramento apoiem diretamente os objetivos de negócios.
- Eficiência: priorize os recursos de monitoramento e a atenção nas métricas que importam.
- Proatividade: reconheça e resolva os problemas antes que eles tenham implicações comerciais mais amplas.

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

Orientação para implementação

Para definir com eficácia os KPIs da workload:

1. Comece com os resultados comerciais: Antes de mergulhar nas métricas, entenda o resultado comercial desejado. É aumento de vendas, maior engajamento do usuário ou tempos de resposta mais rápidos?
2. Correlacione métricas técnicas com objetivos de negócios: Nem todas as métricas técnicas têm um impacto direto nos resultados comerciais. Identifique aquelas que têm, mas geralmente é mais fácil identificar um problema usando um KPI comercial.
3. Use o Amazon CloudWatch: Empregue o CloudWatch para definir e monitorar métricas que representam seus KPIs.
4. Revise e atualize regularmente os KPIs: À medida que sua workload e seus negócios evoluem, mantenha seus KPIs relevantes.
5. Envolva as partes interessadas: Envolva as equipes técnicas e comerciais na definição e revisão dos KPIs.

Nível de esforço do plano de implementação: médio

Recursos

Práticas recomendadas relacionadas:

- the section called “OPS04-BP02 Implementar a telemetria de aplicações”
- the section called “OPS04-BP03 Implementar a telemetria da experiência do usuário”
- the section called “OPS04-BP04 Implementar a telemetria de dependências”
- the section called “OPS04-BP05 Implementar rastreamento distribuído”

Documentos relacionados:

- AWS Observability Best Practices (Práticas recomendadas de observabilidade da AWS)
- Guia do usuário do CloudWatch
- AWS Observability Skill Builder Course (Curso de desenvolvimento de habilidades de observabilidade da AWS)

Vídeos relacionados:

- Developing an observability strategy (Desenvolvimento de uma estratégia de observabilidade)

Exemplos relacionados:

- Um workshop de observabilidade

OPS04-BP02 Implementar a telemetria de aplicações

A telemetria de aplicações serve como base para a observabilidade da workload. É fundamental emitir uma telemetria que ofereça informações práticas sobre o estado de sua aplicação e a obtenção de resultados técnicos e comerciais. Da solução de problemas à medição do impacto de um novo recurso ou à garantia do alinhamento com os indicadores-chave de performance (KPIs) de negócios, a telemetria de aplicações informa a maneira como você cria, opera e desenvolve sua workload.

Métricas, logs e rastreamentos formam os três pilares principais da observabilidade. Eles servem como ferramentas de diagnóstico que descrevem o estado de sua aplicação. Com o tempo, eles

auxiliam na criação de linhas de base e na identificação de anomalias. No entanto, para garantir o alinhamento entre as atividades de monitoramento e os objetivos de negócios, é fundamental definir e monitorar os KPIs. Os KPIs de negócios geralmente facilitam a identificação de problemas em comparação com métricas técnicas isoladas.

Outros tipos de telemetria, como monitoramento de usuários reais (RUM) e transações sintéticas, complementam essas fontes de dados primárias. O RUM oferece informações sobre as interações do usuário em tempo real, enquanto as transações sintéticas simulam possíveis comportamentos do usuário, ajudando a detectar gargalos antes que usuários reais os encontrem.

Resultado desejado: Obtenha insights acionáveis sobre o desempenho de sua workload. Esses insights permitem que você tome decisões proativas sobre otimização de desempenho, obtenha maior estabilidade da workload, simplifique os processos de CI/CD e utilize recursos de forma eficaz.

Antipadrões comuns:

- Observabilidade incompleta: negligência da incorporação da observabilidade em todas as camadas da workload, resultando em pontos cegos que podem obscurecer insights vitais sobre desempenho e comportamento do sistema.
- Visualização fragmentada dos dados: quando os dados estão espalhados por várias ferramentas e sistemas, torna-se difícil manter uma visão holística da integridade e do desempenho de sua workload.
- Problemas relatados pelo usuário: um sinal de que falta a detecção proativa de problemas por meio da telemetria e do monitoramento de KPI de negócios.

Benefícios de estabelecer esta prática recomendada:

- Tomada de decisão informada: com insights de telemetria e KPIs de negócios, você pode tomar decisões baseadas em dados.
- Eficiência operacional aprimorada: a utilização de recursos baseada em dados leva à economia de custos.
- Estabilidade aprimorada da workload: detecção e resolução de problemas mais rápidas, levando a um melhor tempo de atividade.
- Processos simplificados de CI/CD: os insights dos dados de telemetria facilitam o refinamento dos processos e a entrega confiável do código.

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

Orientação para implementação

Para implementar a telemetria de aplicações para sua workload, use serviços da AWS, como o Amazon CloudWatch e o AWS X-Ray. O Amazon CloudWatch fornece um conjunto abrangente de ferramentas de monitoramento, permitindo que você observe seus recursos e aplicações em ambientes da AWS e on-premises. Ele coleta, rastreia e analisa métricas, consolida e monitora dados de log e responde às mudanças em seus recursos, aprimorando sua compreensão de como a workload opera. Em conjunto, o AWS X-Ray permite rastrear, analisar e depurar suas aplicações, oferecendo uma compreensão profunda do comportamento da sua workload. Com recursos como mapas de serviços, distribuições de latência e cronogramas de rastreamento, o X-Ray fornece informações sobre o desempenho da workload e os gargalos que a afetam.

Etapas da implementação

1. Identifique quais dados coletar: Garanta as métricas, os logs e os rastreamentos essenciais que ofereceriam informações substanciais sobre a integridade, o desempenho e o comportamento de sua workload.
2. Implemente o agente do CloudWatch : O agente do CloudWatch é fundamental na aquisição de métricas do sistema e da aplicação e de logs de sua workload e de sua infraestrutura subjacente. O agente do CloudWatch também pode ser usado para coletar OpenTelemetry ou rastreamentos do X-Ray e enviá-los para o X-Ray.
3. Defina e monitore os KPIs de negócios: Estabeleça métricas personalizadas que se alinham com os seus resultados empresariais.
4. Instrumente sua aplicação com o AWS X-Ray: além de implantar o agente do CloudWatch, é fundamental instrumentar sua aplicação para emitir dados de rastreamento. Esse processo pode fornecer mais informações sobre o comportamento e o desempenho da sua workload.
5. Padronize a coleta de dados em sua aplicação: Padronize as práticas de coleta de dados em toda a sua aplicação. A uniformidade ajuda a correlacionar e analisar dados, fornecendo uma visão abrangente do comportamento de sua aplicação.
6. Analise e aja com base nos dados: Depois que a coleta e a normalização de dados estiverem em vigor, use o Amazon CloudWatch para análise de métricas e logs e o AWS X-Ray para análise de rastreamentos. Essa análise pode gerar informações cruciais sobre a integridade, o desempenho e o comportamento de sua workload, orientando o processo de tomada de decisão.

Nível de esforço do plano de implementação: Alto

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar os indicadores-chave de performance
- OPS04-BP03 Implementar a telemetria da experiência do usuário
- OPS04-BP04 Implementar a telemetria de dependências
- OPS04-BP05 Implementar rastreamento distribuído

Documentos relacionados:

- AWS Observability Best Practices (Práticas recomendadas de observabilidade da AWS)
- Guia do usuário do CloudWatch
- Guia do desenvolvedor do AWS X-Ray
- Instrumentação de sistemas distribuídos para visibilidade operacional
- AWS Observability Skill Builder Course (Curso de desenvolvimento de habilidades de observabilidade da AWS)
- Quais são as novidades do Amazon CloudWatch?
- Quais são as novidades do AWS X-Ray?

Vídeos relacionados:

- AWS re:Invent 2022 - Observability best practices at Amazon (AWS re:Invent 2022: práticas recomendadas de observabilidade na Amazon)
- AWS re:Invent 2022 - Developing an observability strategy (AWS re:Invent 2022: desenvolvimento de uma estratégia de observabilidade)

Exemplos relacionados:

- Um workshop de observabilidade
- Biblioteca de soluções da AWS: monitoramento de aplicações com Amazon CloudWatch

OPS04-BP03 Implementar a telemetria da experiência do usuário

É essencial obter insights profundos sobre as experiências dos clientes e as interações com sua aplicação. O monitoramento de usuários reais (RUM) e as transações sintéticas servem como ferramentas poderosas para essa finalidade. O RUM fornece dados sobre interações reais do usuário, oferecendo uma perspectiva não filtrada da satisfação do usuário, enquanto as transações sintéticas simulam as interações do usuário, ajudando a detectar possíveis problemas antes mesmo que eles afetem os usuários reais.

Resultado desejado: Uma visão holística da experiência do cliente, detecção proativa de problemas e otimização das interações do usuário para oferecer experiências digitais perfeitas.

Antipadrões comuns:

- Aplicações sem monitoramento de usuários reais (RUM):
  - Detecção atrasada de problemas: sem o RUM, talvez você não fique ciente dos gargalos ou problemas de desempenho até que os usuários reclamem. Essa abordagem reativa pode levar à insatisfação do cliente.
  - Falta de insights sobre a experiência do usuário: não usar o RUM significa perder dados cruciais que mostram como usuários reais interagem com sua aplicação, limitando sua capacidade de otimizar a experiência do usuário.
- Aplicações sem transações sintéticas:
  - Casos extremos perdidos: transações sintéticas ajudam você a testar caminhos e funções que podem não ser usados com frequência por usuários comuns, mas são essenciais para determinadas funções de negócios. Sem eles, esses caminhos podem ter problemas de funcionamento e passar despercebidos.
  - Verificação de problemas quando a aplicação não está sendo usada: testes sintéticos regulares podem simular momentos em que usuários reais não estão interagindo ativamente com sua aplicação, garantindo que o sistema sempre funcione corretamente.

Benefícios de estabelecer esta prática recomendada:

- Detecção proativa de problemas: identifique e resolva possíveis problemas antes que eles afetem usuários reais.
- Experiência otimizada do usuário: o feedback contínuo do RUM ajuda a refinar e aprimorar a experiência geral do usuário.

- Informações sobre o desempenho do dispositivo e do navegador: entenda o desempenho da sua aplicação em vários dispositivos e navegadores, permitindo uma maior otimização.
- Fluxos de trabalho de negócios validados: transações sintéticas regulares garantem que as principais funcionalidades e os caminhos críticos permaneçam operacionais e eficientes.
- Desempenho aprimorado da aplicação: utilize as informações coletadas de dados reais do usuário para melhorar a capacidade de resposta e a confiabilidade da aplicação.

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

Orientação para implementação

Para aproveitar o RUM e as transações sintéticas para a telemetria da atividade do usuário, a AWS oferece serviços como Amazon CloudWatch RUM e Amazon CloudWatch Synthetics. Métricas, logs e rastreamentos, juntamente com dados de atividades do usuário, fornecem uma visão abrangente do estado operacional da aplicação e da experiência do usuário.

Etapas da implementação

1. Implemente o Amazon CloudWatch RUM: integre sua aplicação ao CloudWatch RUM para coletar, analisar e apresentar dados reais do usuário.
   a. Use a biblioteca em JavaScript do CloudWatch para integrar o RUM à sua aplicação.
   b. Configure painéis para visualizar e monitorar dados reais do usuário.
2. Configuração do CloudWatch Synthetics: crie canários ou rotinas com script que simulem as interações do usuário com sua aplicação.
   a. Defina fluxos de trabalho e caminhos de aplicação críticos.
   b. Projete canários usando scripts do CloudWatch para simular as interações do usuário nesses caminhos.
   c. Programe e monitore os canários para serem executados em intervalos específicos, garantindo verificações de desempenho consistentes.
3. Analise e aja com base nos dados: Utilize dados de RUM e transações sintéticas para obter insights e tomar medidas corretivas quando anomalias forem detectadas. Use painéis do CloudWatch e alarmes para se manter informado.

Nível de esforço do plano de implementação: médio

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar os indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS04-BP04 Implementar a telemetria de dependências
- OPS04-BP05 Implementar rastreamento distribuído

Documentos relacionados:

- Guia do Amazon CloudWatch RUM
- Guia do Amazon CloudWatch Synthetics

Vídeos relacionados:

- Optimize applications through end user insights with Amazon CloudWatch RUM (Otimização de aplicações por meio de insights sobre o usuário final com o Amazon CloudWatch RUM)
- AWS on Air ft. Real-User Monitoring for Amazon CloudWatch (Monitoramento de usuários reais do Amazon CloudWatch)

Exemplos relacionados:

- Um workshop de observabilidade
- Repositório Git do Amazon CloudWatch RUM Web Client
- Uso do Amazon CloudWatch Synthetics para medir o tempo de carregamento da página

OPS04-BP04 Implementar a telemetria de dependências

A telemetria de dependências é essencial para monitorar a integridade e a performance dos serviços e componentes externos dos quais a workload depende. Ela fornece informações valiosas sobre acessibilidade, tempos limite e outros eventos críticos relacionados a dependências, como DNS, bancos de dados ou APIs de terceiros. Ao instrumentar sua aplicação para emitir métricas, logs e rastreamentos sobre essas dependências, você obtém uma compreensão mais clara dos possíveis gargalos, problemas de performance ou falhas que podem afetar a workload.

Resultado desejado: As dependências das quais a workload depende estão funcionando conforme o esperado, permitindo que você resolva problemas de forma proativa e garanta a performance ideal da workload.

Antipadrões comuns:

- Negligenciar as dependências externas: focar apenas nas métricas internas da aplicação e negligenciar as métricas relacionadas às dependências externas.
- Falta de monitoramento proativo: aguardar o surgimento de problemas em vez de monitorar continuamente a integridade e a performance da dependência.
- Monitoramento em silos: usar várias ferramentas de monitoramento diferentes, o que pode resultar em visualizações fragmentadas e inconsistentes da integridade da dependência.

Benefícios de estabelecer esta prática recomendada:

- Maior confiabilidade da workload: garantindo que as dependências externas estejam consistentemente disponíveis e tenham uma performance ideal.
- Detecção e resolução mais rápidas de problemas: identificação e resolução proativa de problemas com dependências antes que elas afetem a workload.
- Visão abrangente: obtendo uma visão holística dos componentes internos e externos que influenciam a integridade da workload.
- Escalabilidade aprimorada da workload: entendendo os limites de escalabilidade e as características de performance das dependências externas.

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

Orientação para implementação

Implemente a telemetria de dependências começando com a identificação dos serviços, da infraestrutura e dos processos dos quais a workload depende. Quantifique quais são as boas condições quando essas dependências estão funcionando conforme o esperado e determine quais dados são necessários para medi-las. Com essas informações, você pode criar painéis e alertas que fornecem insights para suas equipes de operações sobre o estado dessas dependências. Use ferramentas da AWS para descobrir e quantificar os impactos quando as dependências não puderem desempenhar conforme necessário. Revise continuamente sua estratégia para considerar as mudanças nas prioridades, metas e insights obtidos.

Etapas da implementação

Para implementar a telemetria de dependências de forma eficaz:

1. Identifique dependências externas: colabore com as partes interessadas para identificar as dependências externas das quais a workload depende. As dependências externas podem abranger serviços como bancos de dados externos, APIs de terceiros, rotas de conectividade de rede para outros ambientes e serviços de DNS. O primeiro passo para uma telemetria de dependências eficaz é entender de forma abrangente quais são essas dependências.
2. Desenvolva uma estratégia de monitoramento: depois de ter uma visão clara de suas dependências externas, elabore uma estratégia de monitoramento personalizada para elas. Isso envolve entender a importância de cada dependência, seu comportamento esperado e quaisquer contratos ou metas de nível de serviço associados (SLA ou SLTs). Configure alertas proativos para receber notificações sobre mudanças de status ou desvios de performance.
3. Utilize o Amazon CloudWatch Internet Monitor: ele oferece informações sobre a internet global, ajudando a entender interrupções que podem afetar suas dependências externas.
4. Mantenha-se informado com o AWS Health Dashboard: ele fornece alertas e orientações de correção quando a AWS está enfrentando eventos que podem impactar seus serviços.
5. Instrumente sua aplicação com o AWS X-Ray: o AWS X-Ray fornece informações sobre a performance das aplicações e de suas respectivas dependências subjacentes. Ao rastrear as solicitações do início ao fim, você pode identificar gargalos ou falhas nos serviços ou componentes externos dos quais sua aplicação depende.
6. Use o Amazon DevOps Guru: esse serviço orientado por machine learning identifica problemas operacionais, prevê quando problemas críticos podem ocorrer e recomenda ações específicas a serem tomadas. Ele é inestimável para obter informações sobre dependências e determinar que elas não são a fonte dos problemas operacionais.
7. Monitore regularmente: monitore continuamente métricas e logs relacionados a dependências externas. Configure alertas para comportamento inesperado ou diminuição de performance.
8. Valide após as alterações: sempre que houver uma atualização ou alteração em qualquer uma das dependências externas, valide sua performance e verifique o alinhamento com os requisitos da sua aplicação.

Nível de esforço do plano de implementação: médio

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar os indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS04-BP03 Implementar a telemetria da experiência do usuário
- OPS04-BP05 Implementar rastreamento distribuído

Documentos relacionados:

- O que é o AWS Health?
- Uso do Amazon CloudWatch Internet Monitor
- Guia do desenvolvedor do AWS X-Ray
- Guia do usuário do Amazon DevOps Guru

Vídeos relacionados:

- Visibility into how internet issues impact app performance
- Introduction to Amazon DevOps Guru (Introdução ao Amazon DevOps Guru)

Exemplos relacionados:

- Gaining operational insights with AIOps using Amazon DevOps Guru
- AWS Health Aware

OPS04-BP05 Implementar rastreamento distribuído

O rastreamento distribuído oferece uma maneira de monitorar e visualizar solicitações à medida que elas percorrem vários componentes de um sistema distribuído. Ao capturar dados de rastreamento de várias fontes e analisá-los em uma visão unificada, as equipes podem entender melhor como as solicitações fluem, onde existem gargalos e onde os esforços de otimização devem se concentrar.

Resultado desejado: Obtenha uma visão holística das solicitações que fluem pelo seu sistema distribuído, permitindo depuração precisa, desempenho otimizado e experiências de usuário aprimoradas.

Antipadrões comuns:

- Instrumentação inconsistente: nem todos os serviços em um sistema distribuído são instrumentados para rastreamento.
- Ignorar a latência: foco apenas nos erros e sem considerar a latência ou as degradações graduais do desempenho.

Benefícios de estabelecer esta prática recomendada:

- Visão geral abrangente do sistema: visualização de todo o caminho das solicitações, da entrada à saída.
- Depuração aprimorada: identificação rápida de onde ocorrem falhas ou problemas de desempenho.
- Experiência de usuário aprimorada: monitoramento e otimização com base nos dados reais do usuário, garantindo que o sistema atenda às demandas do mundo real.

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

Orientação para implementação

Comece identificando todos os elementos da workload que exigem instrumentação. Depois que todos os componentes forem contabilizados, utilize ferramentas como o AWS X-Ray e o OpenTelemetry para coletar dados de rastreamento para análise com ferramentas como o X-Ray e o Amazon CloudWatch ServiceLens Map. Faça avaliações regulares com desenvolvedores e complemente essas discussões com ferramentas como Amazon DevOps Guru X-Ray Analytics e X-Ray Insights para ajudar a fazer descobertas mais profundas. Estabeleça alertas a partir de dados de rastreamento para notificar quando os resultados, conforme definido no plano de monitoramento da workload, estiverem em risco.

Etapas da implementação

Para implementar o rastreamento distribuído de forma eficaz:

1. Use o AWS X-Ray: integre o X-Ray à sua aplicação para obter informações sobre seu comportamento, entender seu desempenho e identificar gargalos. Utilize o X-Ray Insights para análise automática de rastreamento.

2. Instrumente seus serviços: verifique se cada serviço, de uma função do AWS Lambda a uma instância do EC2, envia dados de rastreamento. Quanto mais serviços você instrumentar, mais clara será a visão completa.
3. incorpore o monitoramento de usuários reais do CloudWatch e o monitoramento sintético: integre o monitoramento de usuários reais (RUM) e o monitoramento sintético com o X-Ray. Isso permite capturar experiências reais do usuário e simular as interações do usuário para identificar possíveis problemas.
4. Use o agente do CloudWatch: o agente pode enviar rastreamentos a partir do X-Ray ou do OpenTelemetry, aumentando a profundidade dos insights obtidos.
5. Use o Amazon DevOps Guru: o DevOps Guruusa dados do X-Ray, CloudWatch, AWS Config e AWS CloudTrail para fornecer recomendações práticas.
6. Analise os rastreamentos: analise regularmente os dados de rastreamento para discernir padrões, anomalias ou gargalos que possam afetar o desempenho de sua aplicação.
7. Configure alertas: configure os alarmes no CloudWatch para padrões incomuns ou latências estendidas, permitindo o tratamento proativo de problemas.
8. Melhoria contínua: revise sua estratégia de rastreamento à medida que os serviços são adicionados ou modificados para capturar todos os pontos de dados relevantes.

Nível de esforço do plano de implementação: médio

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar os indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS04-BP03 Implementar a telemetria da experiência do usuário
- OPS04-BP04 Implementar a telemetria de dependências

Documentos relacionados:

- Guia do desenvolvedor do AWS X-Ray
- Guia do usuário do agente do Amazon CloudWatch
- Guia do usuário do Amazon DevOps Guru

Vídeos relacionados:

- Use AWS X-Ray Insights (Use o AWS X-Ray Insights)
- AWS on Air ft. Observability: Amazon CloudWatch and AWS X-Ray (Observabilidade: Amazon CloudWatch e AWS X-Ray)

Exemplos relacionados:

- Instrumentação da aplicação com AWS X-Ray

OPERAÇÕES 5. Como reduzir defeitos, facilitar a correção e melhorar o fluxo na produção?

Adote abordagens que melhoram o fluxo de alterações na produção, que acionem refatoração, feedback rápido sobre a qualidade e correção de erros. Isso acelera as alterações benéficas que entram na produção, limita os problemas implantados e alcança 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 valide 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 ativar 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 o AWS CodeCommit para gerenciar código e outros artefatos, como modelos do AWS CloudFormation com controle de versão da sua infraestrutura.

Resultado desejado: Suas equipes colaboram no código. Quando mesclado, o código é consistente e nenhuma alteração é perdida. Os erros são facilmente revertidos por meio do versionamento correto.

Antipadrões comuns:

- Você está desenvolvendo e armazenando seu código na estação de trabalho. Você teve uma falha de armazenamento irrecuperável na estação de trabalho e seu código foi perdido.
- Depois de substituir o código existente pelas alterações, você reinicia a aplicação e ela deixa de ser operável. Não é possível reverter 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 a lista de compras sobre relatório final. Não é possível reverter a alteração e você terá que recriar o relatório.

Benefícios de estabelecer esta prática recomendada: Ao usar recursos de versionamento, você pode reverter facilmente para bons estados conhecidos, versões anteriores e limitar o risco de perda de ativos.

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

Orientação para implementação

Mantenha 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.

Recursos

Práticas recomendadas relacionadas:

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

Documentos relacionados:

- O que é o AWS CodeCommit?

Vídeos relacionados:

- Introduction to AWS CodeCommit (Introdução ao AWS CodeCommit)

OPS05-BP02 Testar e valide as alterações

Cada alteração implantada deve ser testada para evitar erros na produção. A prática recomendada concentra-se em testar alterações do controle de versão na build de artefato. Além das alterações do código da aplicação, o teste deve incluir infraestrutura, configuração, controles de segurança e procedimentos de operações. O teste assume muitas formas, desde testes de unidade à análise dos componentes do software (SCA). Mova os testes mais para a esquerda na integração do software e o processo de entrega resultará em maior certeza da qualidade do artefato.

Sua organização deve desenvolver padrões de teste para todos os artefatos de software. Os testes automatizados reduzem o trabalho e evitam erros de testes manuais. Os testes manuais podem ser necessários em alguns casos. Os desenvolvedores precisam ter acesso aos resultados dos testes automatizados para criar loops de feedback que melhorem a qualidade do software.

Resultado desejado: As alterações do software são testadas antes de serem entregues. Os desenvolvedores têm acesso aos resultados e validações dos testes. Sua organização tem um padrão de testes que se aplica a todas as alterações do software.

Antipadrões comuns:

- Você implanta uma nova alteração do software sem nenhum teste. Ele não é executado na produção, o que ocasiona uma interrupção.
- Novos grupos de segurança são implantados com o AWS CloudFormation sem serem testados em um ambiente de pré-produção. Os grupos de segurança tornam sua aplicação inacessível para seus clientes.
- Um método é modificado, mas não há testes de unidade. O software falha quando é implantado em produção.

Benefícios de estabelecer esta prática recomendada: A taxa de falha de alteração de implantações de software é reduzida. A qualidade do software é aprimorada. Os desenvolvedores aumentaram a conscientização sobre a viabilidade do código deles. As políticas de segurança podem ser distribuídas com confiança para apoiar a conformidade da organização. Alterações da infraestrutura, como atualizações da política de ajuste de escala automático, são testadas com antecedência para atender às necessidades de tráfego.

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

Orientação para implementação

Testes são realizados em todas as alterações, desde o código da aplicação à infraestrutura, como parte de sua prática de integração contínua. Os resultados dos testes são publicados para que os desenvolvedores tenham feedback rápido. Sua organização tem um padrão de testes de que todas as alterações devem ser aprovadas.

Exemplo de clientes

Como parte do pipeline de integração contínua, a AnyCompany Retail realiza alguns tipos de teste em todos os artefatos de software. Eles praticam desenvolvimento orientado a testes para que todo o software tenha testes de unidade. Depois que o artefato é criado, eles executam testes completos. Depois que a primeira etapa de testes é concluída, eles executam uma verificação de segurança da aplicação estática, que procura vulnerabilidades conhecidas. Os desenvolvedores recebem mensagens à medida que cada gate de testes é aprovado. Depois que todos os testes são concluídos, o artefato de software é armazenado em um repositório de artefatos.

Etapas da implementação

1. Trabalhe com partes interessadas em sua organização para desenvolver um padrão de testes para artefatos de software. Em quais testes padrão todos os artefatos devem ser aprovados? Há requisitos de conformidade ou governança que devem ser incluídos na cobertura de testes? Você precisa realizar testes de qualidade de código? Quando os testes são concluídos, quem precisa saber?
   a. A arquitetura de referência de pipeline de implantação da AWS contém uma lista confiável de tipos de testes que podem ser conduzidos em artefatos de software como parte de um pipeline de integração.
2. Instrumente sua aplicação com os testes necessários com base em seu padrão de testes de software. Cada conjunto de testes deve ser concluído em menos de dez minutos. Os testes devem ser executados como parte de um pipeline de integração.

a. O Amazon CodeGuru Reviewer pode testar seu código de aplicação quanto a defeitos.
b. Você pode usar o AWS CodeBuild para realizar testes em artefatos de software.
c. O AWS CodePipeline pode orquestrar seus testes de software em um pipeline.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP01 Usar o controle de versão
- OPS05-BP06 Compartilhar os padrões de design
- OPS05-BP10 Automatizar totalmente a integração e a implantação

Documentos relacionados:

- Adopt a test-driven development approach (Adotar uma abordagem de desenvolvimento orientada a testes)
- Automated AWS CloudFormation Testing Pipeline with TaskCat and CodePipeline (Pipeline de teste do AWS CloudFormation automatizado com TaskCat e CodePipeline)
- Building end-to-end AWS DevSecOps CI/CD pipeline with open source SCA, SAST, and DAST tools (Criar um pipeline de CI/CD completo do AWS DevSecOps e ferramentas DAST)
- Getting started with testing serverless applications (Conceitos básicos de testes de aplicações com tecnologia sem servidor)
- My CI/CD pipeline is my release captain (Meu pipeline de CI/CD é meu capitão de lançamentos)
- Whitepaper: Practicing Continuous Integration and Continuous Delivery on AWS (Praticar a integração e entrega contínuas na AWS)

Vídeos relacionados:

- AWS re:Invent 2020: Testable infrastructure: Integration testing on AWS (AWS re:Invent 2020: infraestrutura testável: teste de integração na AWS)
- AWS Summit ANZ 2021 - Driving a test-first strategy with CDK and test driven development (AWS Summit ANZ 2021: conduzir uma estratégia de primeiro teste com o CDK e desenvolvimento orientado a testes)
- Testing Your Infrastructure as Code with AWS CDK (Testar sua infraestrutura como código com o AWS CDK)

Recursos relacionados:

- AWS Deployment Pipeline Reference Architecture - Application (Arquitetura de referência de pipeline de implantação da AWS: aplicação)
- AWS Kubernetes DevSecOps Pipeline (Pipeline do AWS Kubernetes DevSecOps)
- Workshop sobre política como código: desenvolvimento orientado a testes
- Run unit tests for a Node.js application from GitHub by using AWS CodeBuild (Executar testes de unidade para uma aplicação Node.js do GitHub usando o AWS CodeBuild)
- Use Serverspec for test-driven development of infrastructure code (Usar o Serverspec para desenvolvimento orientado a testes do código de infraestrutura)

Serviços relacionados:

- Amazon CodeGuru Reviewer
- AWS CodeBuild
- AWS CodePipeline

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 desfazer a alteração depois do incidente, eliminando os logs agora desnecessários e a despesa associada.

Na AWS, você pode usar o AWS Config para monitorar continuamente as configurações de seus recursos da AWS entre contas e Regiões. Isso ajuda a 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 pacotes de conformidade do AWS Config.

Se tiver configurações dinâmicas em suas aplicações executadas em instâncias do Amazon EC2, AWS Lambda, contêineres, aplicativos móveis ou dispositivos de IoT, você pode usar o AWS AppConfig para configurá-las, validá-las, implantá-las e monitorá-las em seus ambientes.

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

Resultado desejado: Você configura, valida e implanta como parte de seu pipeline de integração contínua e entrega contínua (CI/CD). Você monitora para validar se as configurações estão corretas. Isso minimiza qualquer impacto para usuários finais e clientes.

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.
- Você coloca uma configuração de pré-produção em produção por meio de CI/CD sem validação. Você expõe usuários e clientes a dados e serviços incorretos.

Benefícios de estabelecer esta 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. Os sistemas de gerenciamento de configuração fornecem garantias com relação aos requisitos regulatórios, de conformidade e de governança.

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

Orientação para implementação

Os sistemas de gerenciamento de configuração são usados para rastrear e implementar alterações nas configurações de aplicações e ambientes. Os sistemas de gerenciamento de configuração

também são usados para reduzir erros causados por processos manuais, tornar as alterações de configuração repetíveis e auditáveis e reduzir o nível de esforço.

Etapas da implementação

1. Identifique os proprietários da configuração.
   a. Informe os proprietários das configurações sobre quaisquer necessidades regulatórias, de conformidade ou de controle.
2. Identifique os itens de configuração e os resultados.
   a. Os itens de configuração são todas as configurações de aplicações e ambientes afetadas por uma implantação em seu pipeline de CI/CD.
   b. Os resultados incluem critérios de sucesso, validação e o que monitorar.
3. Selecione ferramentas para gerenciamento de configuração com base nos requisitos de seus negócios e no pipeline de entrega.
4. Considere implantações ponderadas, como implantações canário, para alterações significativas na configuração, a fim de minimizar o impacto de configurações incorretas.
5. Integre seu gerenciamento de configuração ao seu pipeline de CI/CD.
6. Valide todas as alterações enviadas.

Recursos

Práticas recomendadas relacionadas:

- OPS06-BP01 Preparar-se para alterações malsucedidas
- OPS06-BP02 Testar as implantações
- OPS06-BP03 Utilizar estratégias de implantação seguras
- OPS06-BP04 Automatizar os testes e a reversão

Documentos relacionados:

- AWS Control Tower
- AWS Landing Zone Accelerator (Acelerador de zona de pouso da AWS)
- AWS Config
- O que é o AWS Config?
- AWS AppConfig

- O que é o AWS CloudFormation?
- Ferramentas de desenvolvedor da AWS

Vídeos relacionados:

- AWS re:Invent 2022 - Proactive governance and compliance for AWS workloads (AWS re:Invent 2022: governança e conformidade proativas para workloads da AWS)
- AWS re:Invent 2020: Achieve compliance as code using AWS Config (AWS re:Invent 2020: alcance a conformidade como código usando o AWS Config)
- Manage and Deploy Application Configurations with AWS AppConfig (Gerencie e implante configurações de aplicações com o AWS AppConfig)

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

Use 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 de desenvolvedor da AWS (por exemplo, AWS CodeCommit, AWS CodeBuild, o AWS CodePipeline, o AWS CodeDeploye o AWS CodeStar).

Resultado desejado: Seus sistemas de gerenciamento de compilação e implantação oferecem suporte ao sistema de integração contínua (CI/CD) de sua organização, que fornece recursos para automatizar implementações seguras com as configurações corretas.

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 log locais indicam que a falha ocorreu devido à ausência de dependências.
- Você cria a aplicação com êxito com os novos recursos em seu ambiente de desenvolvimento e fornece o código à garantia de qualidade (QA). Ele falha no QA porque não há ativos estáticos.
- Na sexta-feira, após muito esforço, você consegue criar a aplicação 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 de estabelecer esta 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 risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Os sistemas de gerenciamento de compilação e implantação são usados para rastrear e implementar mudanças, reduzir erros causados por processos manuais e reduzir o nível de esforço necessário para implantações seguras. Automatize totalmente o pipeline de integração e implantação desde o check-in do código até a compilação, o teste, a implantação e a validação. Isso reduz o tempo de espera, diminui os custos, incentiva o aumento da frequência de mudanças, reduz o nível de esforço e aumenta a colaboração.

Etapas da implementação

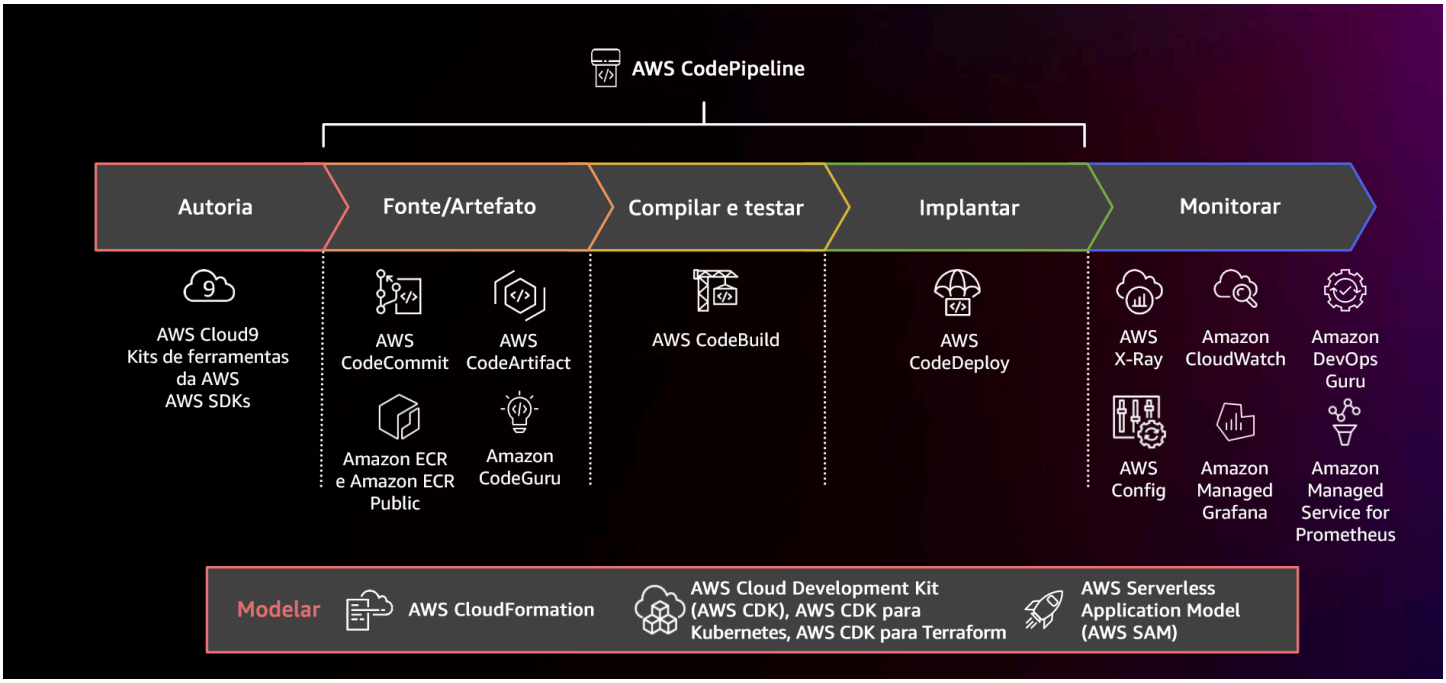


Diagrama mostrando um pipeline de CI/CD usando o AWS CodePipeline e serviços relacionados

1. Use o AWS CodeCommit para controlar versões, armazenar e gerenciar ativos (como documentos, código-fonte e arquivos binários).
2. Use o CodeBuild para compilar seu código-fonte, executar testes de unidade e produzir artefatos prontos para serem implantados.
3. Use CodeDeploy como serviço de implantação que automatiza as implantações de aplicações para Amazon EC2 instâncias, instâncias on-premises, funções do AWS Lambda sem servidorou Amazon ECS.
4. Monitore suas implantações.

Recursos

Práticas recomendadas relacionadas:

- OPS06-BP04 Automatizar os testes e a reversão

Documentos relacionados:

- Ferramentas de desenvolvedor da AWS
- O que é o AWS CodeCommit?
- O que é o AWS CodeBuild?
- AWS CodeBuild
- O que é o AWS CodeDeploy?

Vídeos relacionados:

- AWS re:Invent 2022 - AWS Well-Architected best practices for DevOps on AWS (AWS re:Invent 2022: práticas recomendadas do AWS Well-Architected para DevOps 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, escalar e facilitar a realização de patches.

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.

O Amazon EC2 Image Builder fornece pipelines para atualizar as imagens de máquina. Como parte do gerenciamento de patches, considere Amazon Machine Images (AMIs) usando um pipeline de imagens AMI ou imagens de contêiner com um pipeline de imagens do Docker, enquanto o AWS Lambda fornece padrões para tempos de execução personalizados e bibliotecas adicionais para remover vulnerabilidades.

Você deve gerenciar atualizações em Amazon Machine Images para imagens do Linux ou Windows Server usando o Amazon EC2 Image Builder. Você pode usar o Amazon Elastic Container Registry (Amazon ECR) com seu pipeline existente para gerenciar imagens do Amazon ECS e gerenciar imagens do Amazon EKS. O Lambda inclui recursos de gerenciamento de versão.

A aplicação de patches não deve ser realizada em sistemas de produção sem antes testá-los em um ambiente seguro. Os patches só deverão ser aplicados se forem compatíveis com um resultado operacional ou comercial. Na 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 Systems Manager Maintenance Windows.

Resultado desejado: suas imagens de AMI e contêiner foram corrigidas, atualizadas e estão prontas para o lançamento. Você pode acompanhar o status de todas as imagens implantadas e conhecer a conformidade do patch. É possível emitir relatórios do status atual e ter um processo para atender às suas necessidades de conformidade.

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 a workload.
- Você aplica patches nos ambientes do desenvolvedor automaticamente, sem notificar os desenvolvedores. Você recebe várias reclamações dos desenvolvedores, dizendo que o ambiente deles não está funcionando conforme o esperado.
- Você não aplicou 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.
- Você é notificado sobre uma vulnerabilidade de dia zero que exige uma correção de emergência e precisa fazer isso em todos os seus ambientes manualmente.

Benefícios de estabelecer esta prática recomendada: Ao estabelecer um processo de gerenciamento de patches, incluindo seus critérios de aplicação de patches e metodologia para distribuição em seus ambientes, você pode escalar e gerar relatórios sobre os níveis de patch. Isso fornece garantias sobre a aplicação de patches de segurança e garante uma visibilidade clara do status das correções conhecidas em vigor. Isso permite a adoção de recursos e capacidades desejados, a remoção rápida de problemas e a conformidade contínua com a governança. Implemente sistemas de gerenciamento de patches e automação para reduzir o nível de esforço na implantação de patches e limitar erros causados por processos manuais.

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

Orientação para implementação

Aplique patches nos sistemas para corrigir problemas, obter os recursos ou capacidades desejados e permanecer em conformidade com a política de governança e 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 na aplicação de patches, reduzir erros causados por processos manuais e reduzir o nível de esforço para corrigir.

Etapas da implementação

Para Amazon EC2 Image Builder:

1. Usando o Amazon EC2 Image Builder, especifique os detalhes do pipeline:
    a. Crie um pipeline de imagens e dê um nome a ele
    b. Defina o cronograma e o fuso horário do pipeline
    c. Configure todas as dependências
2. Escolha uma fórmula:
    a. Selecione a fórmula existente ou crie uma nova.
    b. Selecione o tipo de imagem.

   c. Nomeie e crie a versão da sua fórmula
   d. Selecione sua imagem base
   e. Adicione componentes de compilação e adicione ao registro de destino
3. Opcional: defina sua configuração de infraestrutura.
4. Opcional: defina as configurações.
5. Revise as configurações.
6. Mantenha a higiene da fórmula regularmente.

Para o Systems Manager Patch Manager:

1. Crie uma lista de referência de patches.
2. Selecione um método de operações de definição de caminho.
3. Habilite relatórios e escaneamentos de conformidade.

Recursos

Práticas recomendadas relacionadas:

- OPS06-BP04 Automatizar os testes e a reversão

Documentos relacionados:

- O que é o Amazon EC2 Image Builder
- Crie um pipeline de imagens usando o Amazon EC2 Image Builder
- Crie um pipeline de imagens de contêiner
- AWS Systems Manager Patch Manager
- Trabalhar com o Patch Manager
- Trabalhar com relatórios de conformidade de patches
- Ferramentas de desenvolvedor da AWS

Vídeos relacionados:

- CI/CD para aplicações de tecnologia sem servidor na AWS
- Projeto com Ops em mente

Exemplos relacionados:

- Laboratórios do Well-Architected: Gerenciamento de inventário e patches
- Tutoriais do AWS Systems Manager Patch Manager

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. Documente-as e mantenha-as atualizadas à medida que sua arquitetura evolui. Se forem aplicados padrões compartilhados na sua organização, será fundamental que haja mecanismos para solicitar adições, alterações e exceções para os padrões. Sem essa opção, os padrões se tornam uma restrição à inovação.

Resultado desejado: padrões de design são compartilhados entre as equipes nas organizações. Eles são documentados e mantidos atualizados de acordo com a evolução das práticas recomendadas.

Antipadrões comuns:

- Cada uma das duas equipes de desenvolvimento criou um serviço de autenticação de usuários. Os usuários devem manter um conjunto separado de credenciais para cada parte do sistema que desejam acessar.
- Cada equipe gerencia sua própria infraestrutura. Um novo requisito de conformidade força uma alteração na infraestrutura e cada equipe o implementa de maneira diferente.

Benefícios de estabelecer esta prática recomendada: o uso dos padrões compartilhados apoia a adoção das práticas recomendadas e maximiza os benefícios dos esforços de desenvolvimento. A documentação e atualização dos padrões de design mantém a organização atualizada com relação às práticas recomendadas e aos requisitos de segurança e conformidade.

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

Orientação para implementação

Compartilhe as práticas recomendadas, os padrões de design, as listas de verificação, os procedimentos operacionais, as orientações e os requisitos de governança entre equipes. Tenha procedimentos para solicitar alterações, adições e exceções para padrões de design a fim de apoiar a melhoria e a inovação. As equipes devem estar cientes do conteúdo publicado. Tenha um mecanismo para manter os padrões de design atualizados à medida que surgem novas práticas recomendadas.

Exemplo de clientes

A AnyCompany Retail tem uma equipe de arquitetura multifuncional que cria padrões de arquitetura de software. Essa equipe cria a arquitetura com conformidade e governança integradas. As equipes que adotam esses padrões compartilhados recebem os benefícios de ter a conformidade e governança integradas. Elas podem criar rapidamente com base no padrão de design. A equipe de arquitetura se reúne trimestralmente para avaliar os padrões de arquitetura e atualizá-los, se necessário.

Etapas da implementação

1. Identifique uma equipe multifuncional que seja responsável pelo desenvolvimento e pela atualização dos padrões de design. Essa equipe deverá trabalhar com as partes interessadas na organização para desenvolver os padrões de design, os procedimentos operacionais, as listas de verificações, as orientações e os requisitos de governança. Documente os padrões de design e compartilhe-os na organização.
   a. O AWS Service Catalog pode ser usado para criar portfólios representando os padrões de design usando infraestrutura como código. É possível compartilhar portfólios entre contas.
2. Tenha um mecanismo em vigor para manter os padrões de design atualizados à medida que novas práticas recomendadas são identificadas.
3. Se os padrões de design forem aplicados centralmente, tenha um processo para solicitar alterações, atualizações e isenções.

Nível de esforço do plano de implementação: médio. O desenvolvimento de um processo para criar e compartilhar padrões de design pode exigir coordenação e cooperação com as partes interessadas na organização.

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP03 Avaliar os requisitos de governança – Os requisitos de governança influenciam os padrões de design.
- OPS01-BP04 Avaliar os requisitos de conformidade – A conformidade é um fator fundamental na criação dos padrões de design.
- OPS07-BP02 Garantir uma análise consistente da prontidão operacional – As listas de verificação de prontidão operacional são um mecanismo para implementar os padrões de design ao projetar a workload.

- OPS11-BP01 Ter um processo para a melhoria contínua – A atualização dos padrões de design faz parte da melhoria contínua.
- OPS11-BP04 Realizar o gerenciamento de conhecimento – Como parte da sua prática de gerenciamento de conhecimento, documente e compartilhe os padrões de design.

Documentos relacionados:

- Automatizar AWS Backups com AWS Service Catalog
- Conta do AWS Service Catalog aprimorada para o setor
- Como o Expedia Group criou uma oferta de banco de dados como serviço (DBaaS) usando o AWS Service Catalog
- Manter a visibilidade sobre o uso dos padrões de arquitetura de nuvem
- Simplifique o compartilhamento de seus portfólios do AWS Service Catalog em uma configuração do AWS Organizations

Vídeos relacionados:

- AWS Service Catalog – conceitos básicos
- AWS re:Invent 2020: gerencie seus portfólios do AWS Service Catalog como especialista

Exemplos relacionados:

- Arquitetura de referência do AWS Service Catalog
- Workshop do AWS Service Catalog

Serviços relacionados:

- AWS Service Catalog

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, adoção de padrões e programação de pares. Incorpore essas práticas em seu processo de entrega e integração contínua.

Resultado desejado: Sua organização usa práticas recomendadas como análises de código ou programação de pares para melhorar a qualidade do código. Os desenvolvedores e operadores adotam práticas recomendadas de qualidade do código como parte do ciclo de vida de desenvolvimento de software.

Antipadrões comuns:

- Você confirma o código para a ramificação principal da aplicação sem uma análise de código. A alteração é implantada automaticamente na produção e causa uma interrupção.
- Uma nova aplicação é desenvolvida sem nenhum teste de integração, completo ou de unidade. Não há como testar a aplicação antes da implantação.
- Sua equipe faz alterações manuais na produção para solucionar os defeitos. As alterações não passam por testes ou análises de código e não são capturadas nem registradas por processos contínuos de entrega e integração.

Benefícios de estabelecer esta prática recomendada: Ao adotar práticas para melhorar a qualidade do código, é possível reduzir os problemas surgidos na produção. A qualidade do código aumenta com o uso de práticas recomendadas como programação de pares e análises de código.

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

Orientação para implementação

Implemente práticas para melhorar a qualidade do código visando a minimizar os defeitos antes que eles sejam implantados. Use práticas como desenvolvimento orientado por testes, análises de código e programação de pares para aumentar a qualidade do desenvolvimento.

Exemplo de clientes

A AnyCompany Retail adota várias práticas para melhorar a qualidade do código. O desenvolvimento orientado por testes foi adotado como padrão para escrever aplicações. Para alguns recursos novos, os desenvolvedores fazem a programação em pares durante um sprint. Cada pull request passa por uma análise de código feita por um desenvolvedor sênior antes de ser integrada e implantada.

Etapas da implementação

1. Adote práticas de qualidade de código como desenvolvimento orientado por testes, análises de código e programação de pares em seu processo de entrega e integração contínua. Use essas técnicas para melhorar a qualidade do software.

a. O Amazon CodeGuru Reviewer pode fornecer recomendações de programação para código Java e Python usando machine learning.

b. Você pode criar ambientes de desenvolvimento compartilhados com o AWS Cloud9, onde você pode colaborar no desenvolvimento de código.

Nível de esforço do plano de implementação: Médio. Há muitas maneiras de implementar essa prática recomendada, mas pode ser difícil garantir a adesão organizacional.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP06 Compartilhar os padrões de design – É possível compartilhar padrões de design como parte de sua prática de qualidade de código.

Documentos relacionados:

- Agile Software Guide (Guia do software Agile)
- My CI/CD pipeline is my release captain (Meu pipeline de CI/CD é meu capitão de lançamentos)
- Análises de código automatizadas com o Amazon CodeGuru Reviewer
- Adopt a test-driven development approach (Adotar uma abordagem de desenvolvimento orientada a testes)
- How DevFactory builds better applications with Amazon CodeGuru (Como o DevFactory cria melhores aplicações com o Amazon CodeGuru)
- On Pair Programming (Sobre a programação de pares)
- RENGA Inc. automates code reviews with Amazon CodeGuru (RENGA Inc. automatiza análises de código com o Amazon CodeGuru)
- The Art of Agile Development: Test-Driven Development (A arte do desenvolvimento ágil: desenvolvimento orientado por testes)
- Why code reviews matter (and actually save time!) (Por que as análises de código são importantes (e economizam tempo!)

Vídeos relacionados:

- AWS re:Invent 2020: Continuous improvement of code quality with Amazon CodeGuru (AWS re:Invent 2020: melhoria contínua da qualidade do código com o Amazon CodeGuru)
- AWS Summit ANZ 2021 - Driving a test-first strategy with CDK and test driven development (AWS Summit ANZ 2021: conduzir uma estratégia de primeiro teste com o CDK e desenvolvimento orientado a testes)

Serviços relacionados:

- Amazon CodeGuru Reviewer
- Amazon CodeGuru Profiler
- AWS Cloud9,

OPS05-BP08 Usar vários ambientes

Use vários ambientes para experimentar, desenvolver e testar a workload. Use níveis crescentes de controles à medida que os ambientes se aproximam da produção para adquirir confiança de que sua workload operará conforme pretendido quando implantada.

Resultado desejado: Você tem vários ambientes que refletem suas necessidades de conformidade e governança. Você testa e promove o código por meio de ambientes em seu caminho para a produção.

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 indisponível a aplicação para seus usuários.
- Você está operando um serviço multilocatário e não consegue oferecer suporte a uma solicitação do cliente para um ambiente dedicado.

- Nem sempre você testa, mas, quando o faz, o teste acontece em seu ambiente de produção.
- Você acredita que a simplicidade de um único ambiente substitui o escopo do impacto das alterações dentro do ambiente.

Benefícios de estabelecer esta prática recomendada: Você pode oferecer suporte a vários ambientes simultâneos de desenvolvimento, teste e produção, sem criar conflitos entre desenvolvedores ou comunidades de usuários.

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

Orientação para implementação

Use vários ambientes e forneça aos desenvolvedores ambientes de sandbox com controles minimizados para permitir a experimentação. Forneça ambientes de desenvolvimento individuais para ajudar 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 melhorar resultados válidos.

Recursos

Documentos relacionados:

- Programador de instâncias na AWS
- 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. Quando usadas em conjunto com sistemas de gerenciamento de mudanças, sistemas de gerenciamento de configuração e sistemas de compilação e entrega, mudanças frequentes, pequenas e reversíveis reduzem o escopo e o impacto de uma mudança. Isso resulta em solução de problemas mais eficaz e correção mais rápida, com a opção de reverter alterações.

Antipadrões comuns:

- Você implanta uma nova versão de sua aplicação trimestralmente com uma janela de alteração que significa que um serviço principal está desativado.
- Você frequentemente faz alterações no esquema do banco de dados sem rastrear as alterações nos sistemas de gerenciamento.
- Você realiza atualizações manuais no local, substituindo as instalações e configurações existentes e não tem um plano claro de reversão.

Benefícios de estabelecer esta prática recomendada: Os esforços de desenvolvimento são mais rápidos com a implantação frequente de pequenas mudanças. Quando as alterações são pequenas, é muito mais fácil identificar se elas têm consequências indesejadas e são mais fáceis de serem revertidas. Quando as alterações são reversíveis, há menos risco de implementar a alteração à medida que a recuperação é simplificada. O processo de mudança tem um risco reduzido e o impacto de uma alteração malsucedida é reduzido.

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

Orientação para implementação

Use alterações frequentes, pequenas e reversíveis para reduzir o escopo e o impacto de uma alteração. Isso facilita a solução de problemas, ajuda a fazer uma correção mais rápida e oferece a opção de reverter uma alteração. Também aumenta a taxa na qual você pode agregar valor aos negócios.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP03 Usar sistemas de gerenciamento de configuração
- OPS05-BP04 Usar sistemas de gerenciamento de compilação e de implantação
- OPS06-BP04 Automatizar os testes e a reversão

Documentos relacionados:

- Implementing Microservices on AWS (Implementação de microsserviços na AWS)
- Microservices - Observability (Microsserviços: observabilidade)

OPS05-BP10 Automatizar totalmente a integração e a implantação

Automatize a construção, implantação e o teste da workload. 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 consistente de marcação para permitir a identificação dos seus recursos. Identifique seus recursos de organização, contabilidade de custos, controles de acesso pensando na execução de atividades operacionais automatizadas.

Resultado desejado: os desenvolvedores usam ferramentas para entregar códigos e levá-los até a produção. Os desenvolvedores não precisam fazer login no AWS Management Console para fazer atualizações. Há uma trilha de auditoria completa de alterações e configurações, atendendo às necessidades de governança e conformidade. Os processos são repetíveis e padronizados entre as equipes. Os desenvolvedores podem se concentrar no desenvolvimento e na introdução de código, aumentando a produtividade.

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 de código e cada um dos scripts de teste de unidade, você registra seu código para a próxima versão programada.
- 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 código para solicitar aprovação para implantá-lo na produção.
- Como desenvolvedor, você faz login no AWS Management Console para criar um novo ambiente de desenvolvimento usando métodos e sistemas que não são padrão.

Benefícios de estabelecer esta 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, ajudando os membros da equipe a se concentrarem na entrega de valor para a empresa. Você aumenta a velocidade de entrega à medida que avança até a produção.

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

Orientação para implementação

Você usa sistemas de gerenciamento de criação e 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, o teste, a implantação e a validação. Isso reduz o tempo de espera, aumenta a frequência de alterações, reduz o nível de esforço, aumenta a velocidade de entrada no mercado, resulta em maior produtividade e aumenta a segurança do seu código à medida que você o leva até a produção.

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP03 Usar sistemas de gerenciamento de configuração
- OPS05-BP04 Usar sistemas de gerenciamento de compilação e de implantação

Documentos relacionados:

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

Vídeos relacionados:

- AWS re\:Invent 2022 - AWS Well-Architected best practices for DevOps on AWS

OPERAÇÕES 6. Como reduzir os riscos de implantação?

Adote abordagens que forneçam feedback rápido sobre a qualidade e alcancem 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 Preparar-se para alterações malsucedidas
- OPS06-BP02 Testar as implantações
- OPS06-BP03 Utilizar estratégias de implantação seguras
- OPS06-BP04 Automatizar os testes e a reversão

OPS06-BP01 Preparar-se para alterações malsucedidas

Planeje reverter para um bom estado anterior ou realize reparos no ambiente de produção se a implantação causar um resultado indesejado. Ter uma política para estabelecer esse plano ajuda

todas as equipes a desenvolver estratégias para se recuperar de alterações com falha. Alguns exemplos de estratégias são etapas de implantação e reversão, políticas de alteração, sinalizadores de atributos, isolamento de tráfego e mudança de tráfego. Uma única versão pode incluir várias alterações de componentes relacionadas. A estratégia deve fornecer a possibilidade de resistir ou se recuperar de uma falha de qualquer alteração de componente.

Resultado desejado: Você preparou um plano de recuperação detalhado para a alteração, caso ela não tenha êxito. Além disso, você reduziu o tamanho da sua versão para minimizar o impacto potencial em outros componentes da workload. Como resultado, você reduziu o impacto nos negócios ao diminuir o possível tempo de inatividade decorrente de uma alteração malsucedida e aumentou a flexibilidade e a eficiência dos tempos de recuperação.

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 permanecerá inacessível.
- Seus sistemas não são arquitetados de uma forma que permite que sejam atualizados com versões menores. Como resultado, você tem dificuldade em reverter essas alterações em massa durante uma implantação com falha.
- Você não usa a infraestrutura como código (IaC) e foram feitas atualizações manuais nela que resultaram em uma configuração indesejada. Você não consegue rastrear e reverter com eficácia as alterações manuais.
- Como você não mediu o aumento da frequência das implantações, sua equipe não é incentivada a reduzir o tamanho das mudanças e melhorar seus planos de reversão para cada uma delas, gerando mais riscos e maiores taxas de falha.
- Você não mede a duração total de uma interrupção causada por alterações malsucedidas. A equipe não consegue priorizar e melhorar a eficácia do processo de implantação e do plano de recuperação.

Benefícios de estabelecer esta prática recomendada: Ter um plano para se recuperar de mudanças malsucedidas minimiza o tempo médio de recuperação (MTTR) e reduz o impacto nos negócios.

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

Orientação para implementação

A adoção de uma política e prática documentadas e consistentes por parte das equipes de lançamento permitem que a organização planeje o que deve ocorrer se houver mudanças malsucedidas. A política deve permitir a correção em circunstâncias específicas. Seja qual for a situação, um plano de correção antecipada ou reversão deve ser bem documentado e testado antes da implantação na produção em tempo real, a fim de que o tempo necessário para reverter uma alteração seja minimizado.

Etapas da implementação

1. Documente as políticas que exigem que as equipes tenham planos efetivos para reverter as mudanças dentro de um período especificado.
   a. As políticas devem especificar quando uma situação de correção antecipada é permitida.
   b. Exija que um plano de reversão documentado seja acessível a todos os envolvidos.
   c. Especifique os requisitos de reversão (por exemplo, quando for constatado que foram implantadas alterações não autorizadas).
2. Analise o nível de impacto de todas as mudanças relacionadas a cada componente de uma workload.
   a. Permita que alterações repetíveis sejam padronizadas, modeladas e pré-autorizadas se seguirem um fluxo de trabalho consistente que imponha políticas de mudança.
   b. Reduza o impacto potencial de qualquer alteração diminuindo o tamanho dela para que a recuperação leve menos tempo e cause um impacto menor nos negócios.
   c. Garanta que os procedimentos de reversão revertam o código para um bom estado conhecido a fim de evitar incidentes sempre que possível.
3. Integre ferramentas e fluxos de trabalho para aplicar suas políticas de forma programática.
4. Torne os dados sobre as alterações visíveis para outros proprietários da workload a fim de melhorar a velocidade do diagnóstico de qualquer alteração malsucedida que não possa ser revertida.
   a. Avalie o sucesso dessa prática usando dados de mudança visíveis e identifique melhorias iterativas.
5. Use ferramentas de monitoramento para verificar o sucesso ou a falha de uma implantação a fim de acelerar a tomada de decisões sobre a reversão.

6. Meça a duração da interrupção durante uma alteração malsucedida para melhorar continuamente seus planos de recuperação.

Nível de esforço do plano de implementação: médio

Recursos

Práticas recomendadas relacionadas:

- OPS06-BP04 Automatizar os testes e a reversão

Documentos relacionados:

- AWS Builders Library | Ensuring Rollback Safety During Deployments
- Whitepaper da AWS | Change Management in the Cloud

Vídeos relacionados:

- re:Invent 2019 | A abordagem da Amazon para implantação de alta disponibilidade

OPS06-BP02 Testar as implantações

Teste os procedimentos de lançamento na pré-produção usando a mesma configuração de implantação, controles de segurança, etapas e procedimentos da produção. Valide se todas as etapas implantadas foram concluídas conforme o esperado, como inspecionar arquivos, configurações e serviços. Teste ainda mais todas as alterações com testes funcionais, de integração e de carga, além de qualquer monitoramento, como verificações de integridade. Ao fazer esses testes, você pode identificar problemas de implantação com antecedência, podendo planejá-los e mitigá-los antes da produção.

Você pode criar ambientes paralelos temporários para testar cada alteração. Automatize a implantação dos ambientes de teste usando a infraestrutura como código (IaC) para ajudar a reduzir a quantidade de trabalho envolvido e garantir estabilidade, consistência e entrega mais rápida de atributos.

Resultado desejado: A organização adota uma cultura de desenvolvimento orientada a testes que inclui testes de implantações. Isso garante que as equipes se concentrem em oferecer valor

empresarial em vez de gerenciar lançamentos. As equipes são engajadas desde o início após a identificação dos riscos de implantação para determinar o curso apropriado da mitigação.

Antipadrões comuns:

- Durante as versões de produção, implantações não testadas causam problemas frequentes que exigem soluções e encaminhamento.
- Sua versão contém infraestrutura como código (IaC) que atualiza os recursos existentes. Você não tem certeza se a IaC será executada com êxito ou causará impacto nos recursos.
- Você implanta um novo atributo na aplicação. Ele não funciona conforme o esperado e não há visibilidade até ser relatado pelos usuários afetados.
- Você atualiza seus certificados. Você instala acidentalmente os certificados nos componentes errados, o que não é detectado e afeta os visitantes do site porque não é possível estabelecer uma conexão segura.

Benefícios de estabelecer esta prática recomendada: Testes extensivos na pré-produção dos procedimentos de implantação, considerando-se que as mudanças introduzidas por eles minimizam o impacto potencial na produção causado pelas etapas de implantação. Isso aumenta a confiança durante o lançamento da produção e minimiza o suporte operacional sem diminuir a velocidade das alterações que estão sendo entregues.

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

Orientação para implementação

Testar seu processo de implantação é tão importante quanto testar as alterações resultantes da implantação. Isso pode ser realizado testando suas etapas de implantação em um ambiente de pré-produção que se assemelhe o máximo possível à produção. Problemas comuns, como etapas de implantação incompletas ou incorretas, ou configurações incorretas, podem ser detectados como resultado antes da produção. Além disso, você pode testar suas etapas de recuperação.

Exemplo de clientes

Como parte do pipeline de integração e entrega contínuas (CI/CD), a Loja UmaEmpresa executa as etapas definidas necessárias para lançar atualizações de infraestrutura e software para seus clientes em um ambiente semelhante ao de produção. O pipeline é composto por pré-verificações para detectar desvios (detecção de alterações nos recursos executados fora da IaC) nos recursos antes da implantação, bem como validar as ações que a IaC realiza após seu início. Ele valida as

etapas de implantação, como verificar se determinados arquivos e configurações estão em vigor e se os serviços estão em execução e respondendo corretamente às verificações de integridade no host local antes de serem registrados novamente no balanceador de carga. Além disso, todas as alterações sinalizam vários testes automatizados, como testes funcionais, de segurança, de regressão, de integração e de carga.

Etapas para a implementação

1. Execute verificações de pré-instalação para espelhar o ambiente de pré-produção na produção.
   a. Use detecção de desvios para detectar quando os recursos foram alterados fora do AWS CloudFormation.
   b. Use conjuntos de alterações para validar se a intenção da atualização da pilha corresponde às ações que o AWS CloudFormation realiza quando o conjunto de alterações é iniciado.
2. Isso aciona uma etapa de aprovação manual no AWS CodePipeline para autorizar a implantação no ambiente de pré-produção.
3. Use configurações de implantação, como arquivos do AWS CodeDeploy AppSpec para definir as etapas de implantação e validação.
4. Quando aplicável, integre o AWS CodeDeploy a outros serviços da AWS ou integre o AWS CodeDeploy a produtos e serviços de parceiros.
5. Monitore as implantações usando notificações de eventos do Amazon CloudWatch, do AWS CloudTrail e do Amazon SNS.
6. Execute testes automatizados pós-implantação, incluindo testes funcionais, de segurança, regressão, integração e carga.
7. Solução de problemas de implantação.
8. A validação bem-sucedida das etapas acima deve iniciar um fluxo de trabalho de aprovação manual para autorizar a implantação na produção.

Nível de esforço do plano de implementação: Alto

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP02 Testar e valide as alterações

Documentos relacionados:

- AWS Builders' Li| Automating safe, hands-off deployments | Test Deployments
- Whitepaper da AWS Whitepaper | Practicing Continuous Integration and Continuous Delivery on AWS
- A história da Apollo: o mecanismo de implantação da Amazon
- How to test and debug AWS CodeDeploy locally before you ship your code
- Integrar testes de conectividade de rede com implantação de infraestrutura

Vídeos relacionados:

- re:Invent 2020 | Testar software e sistemas na Amazon

Exemplos relacionados:

- Tutorial | Deploy and Amazon ECS service with a validation test

OPS06-BP03 Utilizar estratégias de implantação seguras

Implantações seguras de produção controlam o fluxo de mudanças benéficas com o objetivo de minimizar qualquer impacto percebido dessas alterações para os clientes. Os controles de segurança fornecem mecanismos de inspeção para validar os resultados desejados e limitar o escopo do impacto dos defeitos introduzidos pelas alterações ou das falhas de implantação. As implementações seguras podem incluir estratégias como sinalizadores de atributos e implantações one-box, contínuas (versões canário), imutáveis, de divisão de tráfego e azuis/verdes.

Resultado desejado: sua organização usa um sistema de integração e entrega contínuas (CI/CD) que fornece recursos para automatizar implementações seguras. As equipes devem usar estratégias apropriadas de implantação seguras.

Antipadrões comuns:

- Você implanta uma alteração malsucedida em toda a produção de uma só vez. Como resultado, todos os clientes são afetados simultaneamente.
- Um defeito introduzido em uma implantação simultânea em todos os sistemas requer um lançamento de emergência. A correção para todos os clientes leva vários dias.
- O gerenciamento da versão de produção requer planejamento e participação de várias equipes. Isso restringe sua capacidade de atualizar atributos com frequência para seus clientes.

- Você executa uma implantação mutável modificando os sistemas existentes. Depois de descobrir que a alteração não foi bem-sucedida, você será forçado a modificar os sistemas novamente para restaurar a versão antiga, aumentando o seu tempo de recuperação.

Benefícios de estabelecer esta prática recomendada: implantações automatizadas equilibram a velocidade das implementações com a entrega consistente de mudanças benéficas aos clientes. Limitar o impacto evita falhas de implantação dispendiosas e maximiza a capacidade das equipes de responder às falhas de forma eficiente.

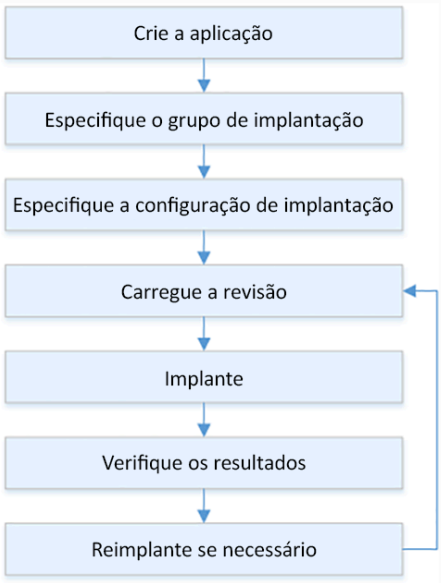
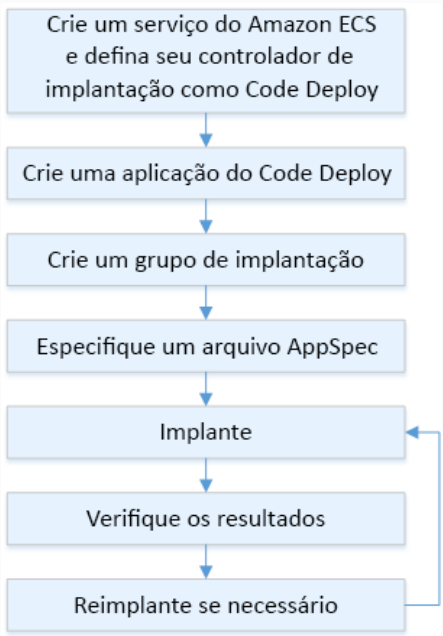
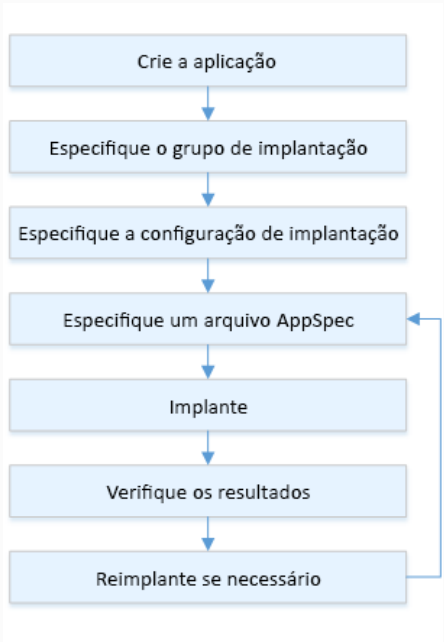
Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Falhas na entrega contínua podem levar à redução da disponibilidade do serviço e a uma experiência ruim para o cliente. Para maximizar a taxa de implantações bem-sucedidas, implemente controles de segurança no processo de lançamento de ponta a ponta para minimizar os erros de implantação e eliminar as falhas.

Exemplo de clientes

A Loja UmaEmpresa tem a missão de alcançar implantações com tempo de inatividade entre mínimo e zero, isto é, sem impacto perceptível para seus usuários durante as implantações. Para fazer isso, a empresa estabeleceu padrões de implantação (consulte o diagrama de fluxo de trabalho a seguir), como implantações azuis/verdes e contínuas. Todas as equipes adotam um ou mais desses padrões no pipeline de CI/CD.

Fluxo de trabalho do CodeDeploy para Amazon EC2	Fluxo de trabalho do CodeDeploy para Amazon ECS	Fluxo de trabalho do CodeDeploy para Lambda
		

Etapas da implementação

1. Use um fluxo de trabalho de aprovação para iniciar a sequência das etapas de implantação na promoção para implantação.
2. Use um sistema de implantação automatizado, como o AWS CodeDeployAs opções de implantação do AWS CodeDeploy incluem implantações locais do EC2/on-premises e implantações azuis/verdes do EC2/on-premises, do AWS Lambda e do Amazon ECS (consulte o diagrama de fluxo de trabalho anterior).
   a. Quando aplicável, integre o AWS CodeDeploy a outros serviços da AWS ou integre o AWS CodeDeploy a produtos e serviços de parceiros.
3. Use implantações azuis/verdes para bancos de dados, como Amazon Aurora e o Amazon RDS.
4. Monitore as implantações usando notificações de eventos do Amazon CloudWatch, do AWS CloudTrail e do Amazon SNS.
5. Realize testes automatizados pós-implantação, incluindo testes funcionais, de segurança, regressão, integração e testes de carga.
6. Solução de problemas de implantação.

Nível de esforço do plano de implementação: médio

Recursos

Práticas recomendadas relacionadas:

- OPS05-BP02 Testar e valide as alterações
- OPS05-BP09 Fazer alterações frequentes, pequenas e reversíveis
- OPS05-BP10 Automatizar totalmente a integração e a implantação

Documentos relacionados:

- AWS Builders Library | Automating safe, hands-off deployments | Production deployments
- AWS Builders Library | My CI/CD pipeline is my release captain | Safe, automatic production releases
- Whitepaper da AWS | Practicing Continuous Integration and Continuous Delivery on AWS | Deployment methods
- Guia do usuário do AWS CodeDeploy
- Working with deployment configurations in AWS CodeDeploy
- Set up an API Gateway canary release deployment
- Amazon ECS Deployment Types
- Fully Managed Blue/Green Deployments in Amazon Aurora and Amazon RDS
- Blue/Green deployments with AWS Elastic Beanstalk

Vídeos relacionados:

- re:Invent 2020 | Sem intervenção manual: como automatizar os pipelines de entrega contínua na Amazon
- re:Invent 2019 | A abordagem da Amazon para implantação de alta disponibilidade

Exemplos relacionados:

- Try a Sample Blue/Green Deployment in AWS CodeDeploy
- Workshop | Buiding CI/CD pipelines for Lambda canary deployments using AWS CDK

- Workshop | Implantação canário e azul/verde para o EKS e o ECS
- Workshop | Criar um pipeline de CI/CD entre contas

OPS06-BP04 Automatizar os testes e a reversão

Para aumentar a velocidade, a confiabilidade e a confiança do seu processo de implantação, tenha uma estratégia para testes automatizados e recursos de reversão em ambientes de pré-produção e produção. Automatize os testes ao implantar na produção para simular interações entre humanos e sistemas que verifiquem as alterações que estão sendo implantadas. Automatize a reversão para voltar rapidamente a um estado anterior em boas condições. A reversão deve ser iniciada automaticamente em condições predefinidas, como quando o resultado desejado da alteração não é alcançado ou quando o teste automatizado falha. A automação dessas duas atividades melhora a taxa de sucesso das implantações, minimiza o tempo de recuperação e reduz o impacto potencial nos negócios.

Resultado desejado: Os testes automatizados e as estratégias de reversão são integrados ao pipeline de integração e entrega contínuas (CI/CD). O monitoramento é capaz de validar seus critérios de sucesso e iniciar a reversão automática em caso de falha. Isso minimiza qualquer impacto para usuários finais e clientes. Por exemplo, quando todos os resultados do teste são satisfatórios, você promove seu código no ambiente de produção em que o teste de regressão automatizado é iniciado, utilizando os mesmos casos de teste. Se os resultados do teste de regressão não corresponderem às expectativas, a reversão automática será iniciada no fluxo de trabalho do pipeline.

Antipadrões comuns:

- Seus sistemas não são arquitetados de uma forma que permite que sejam atualizados com versões menores. Como resultado, você tem dificuldade em reverter essas alterações em massa durante uma implantação com falha.
- O processo de implantação consiste em uma série de etapas manuais. Depois de implantar as alterações na workload, você inicia os testes pós-implantação. Após o teste, você percebe que a workload está inoperante e os clientes estão desconectados. Em seguida, você começa a reverter para a versão anterior. Todas essas etapas manuais atrasam a recuperação geral do sistema e causam um impacto prolongado para os clientes.
- Você passou um tempo desenvolvendo casos de teste automatizados para funcionalidades que não são usadas com frequência na aplicação, minimizando o retorno sobre o investimento no recurso de teste automatizado.

- Sua versão é composta de atualizações de aplicações, infraestrutura, patches e configuração que são independentes umas das outras. No entanto, você tem um único pipeline de CI/CD que fornece todas as alterações de uma só vez. Uma falha em um componente força você a reverter todas as alterações, tornando a reversão complexa e ineficiente.
- A equipe conclui o trabalho de codificação no primeiro sprint e inicia o trabalho no segundo sprint, mas seu plano não incluiu testes até o terceiro sprint. Como resultado, os testes automatizados revelaram defeitos do primeiro sprint que precisavam ser resolvidos antes que o teste dos resultados do segundo sprint pudesse ser iniciado, adiando todo o lançamento e desvalorizando seus testes automatizados.
- Seus casos de teste de regressão automatizados para a versão de produção estão completos, mas você não está monitorando a integridade da workload. Como você não tem visibilidade sobre se o serviço foi reiniciado, você não tem certeza se a reversão é necessária ou se ela já ocorreu.

Benefícios de estabelecer esta prática recomendada: O teste automatizado aumenta a transparência do processo de teste e a capacidade de abranger mais atributos em um período mais curto. Ao testar e validar as mudanças na produção, você pode identificar problemas imediatamente. A melhoria na consistência com ferramentas de teste automatizadas permite uma melhor detecção de defeitos. Ao reverter automaticamente para a versão anterior, o impacto sobre seus clientes é minimizado. A reversão automatizada acaba inspirando mais confiança em seus recursos de implantação ao reduzir o impacto nos negócios. No geral, esses recursos reduzem o tempo de entrega e, ao mesmo tempo, garantem a qualidade.

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

Orientação para implementação

Automatize os testes dos ambientes implantados para confirmar os resultados desejados mais rapidamente. Automatize a reversão para um bom estado anterior conhecido quando os resultados predefinidos não forem alcançados, para minimizar o tempo de recuperação e reduzir os erros causados por processos manuais. Integre ferramentas de teste com seu fluxo de trabalho de pipeline para testar e minimizar as entradas manuais de forma consistente. Priorize a automação de casos de teste, como aqueles que mitigam os maiores riscos e precisam ser testados com frequência a cada alteração. Além disso, automatize a reversão com base em condições específicas predefinidas no plano de teste.

Etapas para a implementação

1. Estabeleça um ciclo de vida de teste para o ciclo de vida de desenvolvimento que defina cada estágio do processo de teste, desde o planejamento dos requisitos até o desenvolvimento do caso de teste, a configuração da ferramenta, o teste automatizado e o encerramento do caso de teste.
   a. Crie uma abordagem de teste específica para workloads com base em sua estratégia geral de teste.
   b. Considere uma estratégia de teste contínuo, quando apropriado, durante o ciclo de vida do desenvolvimento.
2. Selecione ferramentas automatizadas para testes e reversões com base em seus requisitos de negócios e investimentos em pipeline.
3. Decida quais casos de teste você deseja automatizar e quais deverão ser executados manualmente. Eles podem ser definidos com base na prioridade do valor comercial do atributo que está sendo testado. Alinhe todos os membros da equipe a esse plano e verifique a responsabilidade pela realização de testes manuais.
   a. Aplique recursos de teste automatizados a casos de teste específicos que façam sentido para automação, como casos repetíveis ou executados com frequência, aqueles que exigem tarefas repetitivas ou aqueles que são necessários em várias configurações.
   b. Defina scripts de automação de testes, bem como os critérios de sucesso na ferramenta de automação, para que a automação contínua do fluxo de trabalho possa ser iniciada quando casos específicos falharem.
   c. Defina critérios de falha específicos para a reversão automatizada.
4. Priorize a automação de testes para gerar resultados consistentes com o desenvolvimento completo de casos de teste em que a complexidade e a interação humana têm um risco maior de falha.
5. Integre as ferramentas automatizadas de teste e reversão no pipeline de CI/CD.
   a. Desenvolva critérios claros de sucesso para as alterações.
   b. Monitore e observe para detectar esses critérios e reverter automaticamente as alterações quando critérios específicos de reversão forem atendidos.
6. Execute diferentes tipos de teste de produção automatizados, como:
   a. Teste A/B para mostrar resultados em comparação com a versão atual entre dois grupos de teste de usuários.
   b. Teste canário, que permite implantar a alteração em um subconjunto de usuários antes de lançá-la para todos.

   c. Teste de sinalização de atributos, que permite que a sinalização de um único atributo da nova versão seja ativada e desativada de fora da aplicação para que cada novo atributo possa ser validado individualmente.

   d. Teste de regressão para verificar novas funcionalidades com componentes inter-relacionados existentes.

7. Monitore os aspectos operacionais da aplicação, das transações e das interações com outras aplicações e componentes. Desenvolva relatórios para mostrar o sucesso das alterações por workload e identificar quais partes da automação e do fluxo de trabalho podem ser ainda mais otimizadas.

   a. Desenvolva relatórios de resultados de testes que ajudem você a tomar decisões rápidas sobre se os procedimentos de reversão devem ou não ser invocados.

   b. Implemente uma estratégia que permita a reversão automatizada com base em condições de falha predefinidas que resultam de um ou mais de seus métodos de teste.

8. Desenvolva seus casos de teste automatizados para permitir a reutilização em futuras alterações repetíveis.

Nível de esforço do plano de implementação: médio

Recursos

Práticas recomendadas relacionadas:

- OPS06-BP01 Preparar-se para alterações malsucedidas
- OPS06-BP02 Testar as implantações

Documentos relacionados:

- AWS Builders Library | Ensuring rollback safety during deployments
- Redeploy and rollback a deployment with AWS CodeDeploy
- 8 best practices when automating your deployments with AWS CloudFormation

Exemplos relacionados:

- Serverless UI testing using Selenium, AWS Lambda, AWS Fargate (Fargate), and AWS Developer Tools

Vídeos relacionados:

- re:Invent 2020 | Sem intervenção manual: como automatizar os pipelines de entrega contínua na Amazon
- re:Invent 2019 | A abordagem da Amazon para implantação de alta disponibilidade

OPERAÇÕES 7. Como saber se está pronto para oferecer suporte a uma workload?

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-BP06 Viabilizar planos de suporte para workloads de produção

OPS07-BP01 Garantir a capacidade da equipe

Tenha um mecanismo para validar que você tem o número adequado de funcionários treinados para fornecer suporte à workload. Eles devem ter treinamento para a plataforma e os serviços que compõem sua workload. Forneça a eles o conhecimento necessário para operar a workload. É necessário ter o número suficiente de funcionários treinados para fornecer suporte à operação da workload e solucionar os incidentes que ocorrerem. Tenha funcionários suficientes para que seja possível alterná-los durante plantões e férias, a fim de evitar a exaustão.

Resultado desejado:

- Há um número suficiente de funcionários treinados para fornecer suporte à workload quando a workload estiver disponível.
- Você fornece treinamento para seus funcionários sobre software e serviços que compõem a workload.

Antipadrões comuns:

- Implantar uma workload sem membros da equipe treinados para operar a plataforma e os serviços em uso.
- Não ter funcionários suficientes para fornecer suporte à alternâncias de plantão ou folga de funcionários.

Benefícios do estabelecimento desta prática recomendada:

- Ter membros da equipe qualificados possibilita o suporte eficaz da sua carga de trabalho.
- Com um número suficiente de membros na equipe, é possível dar conta da workload e das alternâncias de plantão, reduzindo o risco de exaustão.

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

Orientações para a implementação

Valide se há um número suficiente de funcionários treinados para fornecer suporte à workload. Verifique se você tem membros da equipe suficientes para cobrir as atividades operacionais normais, incluindo alternâncias de plantão.

Exemplo de clientes

A Loja UmaEmpresa garante que as equipes que fornecem suporte à workload estejam completas e treinadas. Há engenheiros suficientes para fornecer suporte a uma alternância de plantão. Os funcionários têm treinamento referente ao software e à plataforma na qual a workload é criada e são incentivados a obter certificações. Há funcionários suficientes para que as pessoas possam tirar folgas enquanto mantêm o suporte à workload e à alternância de plantões.

Etapas da implementação

1. Atribua um número adequado de funcionários para operar e fornecer suporte à workload, incluindo tarefas de plantão.
2. Treine seus funcionários referente ao software e às plataformas que compõem a workload.
    a. O AWS Training and Certification tem uma biblioteca de cursos sobre a AWS. Estão disponíveis cursos pagos e gratuitos, online e presenciais.
    b. A AWS promove eventos e webinars por meio dos quais você aprende com especialistas da AWS.

3. Avalie regularmente o tamanho e as habilidades da equipe à medida que as condições operacionais e a workload mudam. Ajuste o tamanho e as habilidades da equipe para corresponderem aos requisitos operacionais.

Nível de esforço do plano de implementação: alto. Contratar e treinar uma equipe para fornecer suporte a uma workload pode exigir um esforço significativo, mas traz benefícios substanciais de longo prazo.

Recursos

Práticas recomendadas relacionadas:

- OPS11-BP04 Realizar o gerenciamento de conhecimento – Os membros da equipe devem ter as informações necessárias para operar e fornecer suporte à workload. O gerenciamento de conhecimento é fundamental para fornecer isso.

Documentos relacionados:

- Eventos e webinars da AWS
- AWS Training and Certification

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 Preparar-se 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 Realizar 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 Realizar 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

Há processos em vigor para alterações com e sem êxito feitas na workload. Uma estratégia pre-mortem é um exercício em que uma equipe simula uma falha para desenvolver estratégias de mitigação. Use as estratégias pre-mortem para antecipar falhas e criar procedimentos, quando apropriado. Avalie os benefícios e os riscos de implantar alterações na workload. Verifique se todas as alterações estão em conformidade com a governança.

Resultado desejado:

- Você toma decisões embasadas ao implantar alterações na workload.
- As alterações estão em conformidade com a governança.

Antipadrões comuns:

- Implantar uma alteração em nossa workload sem um processo para lidar com uma implantação com falha.
- Fazer alterações no ambiente de produção que estão fora da conformidade com os requisitos de governança.
- Implantar uma nova versão da workload sem estabelecer uma referência para a utilização de recursos.

Benefícios do estabelecimento desta prática recomendada:

- Você está preparado para alterações sem êxito na workload.
- As alterações na workload estão em conformidade com as políticas de governança.

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

Orientações para a implementação

Use estratégias pre-mortem no desenvolvimento de processos para alterações sem êxito. Documente os processos de alterações sem êxito. Garanta que todas as alterações estejam em conformidade com a governança. Avalie os benefícios e os riscos de implantar alterações na workload.

Exemplo de clientes

A Loja UmaEmpresa realiza estratégias pre-mortem regularmente para validar seus processos de alterações sem êxito. Os processos são documentados em uma Wiki compartilhada e atualizados regularmente. Todas as alterações estão em conformidade com os requisitos de governança.

Etapas da implementação

1. Tome decisões embasadas ao implantar alterações na workload. Estabeleça e revise os critérios de uma implantação bem-sucedida. Desenvolva cenários ou critérios que acionariam a reversão de uma alteração. Pondere os benefícios de implantar alterações considerando os riscos de uma alteração sem êxito.
2. Verifique se todas as alterações estão em conformidade com as políticas de governança.
3. Use estratégias pre-mortem para alterações sem êxito e documente as estratégias de migração. Realize um exercício de simulação para modelar uma alteração sem êxito e validar os procedimentos de reversão.

Nível de esforço do plano de implementação: moderado. Implementar uma prática de estratégias pre-mortem requer coordenação e esforço das partes interessadas na organização.

Recursos

Práticas recomendadas relacionadas:

- OPS01-BP03 Avaliar os requisitos de governança – Os requisitos de governança são um fator fundamental para determinar se uma alteração deve ser implementada.
- OPS06-BP01 Preparar-se para alterações malsucedidas – Estabeleça planos para mitigar uma implantação sem êxito e use estratégias pre-mortem para validá-los.
- OPS06-BP02 Testar as implantações – Toda alteração de software deve ser testada adequadamente antes da implantação para reduzir os defeitos na produção.

- OPS07-BP01 Garantir a capacidade da equipe – Ter um número suficiente de funcionários treinados para fornecer suporte à workload é essencial para tomar uma decisão embasada quanto à implantação de uma alteração no sistema.

Documentos relacionados:

- Amazon Web Services: risco e conformidade
- Modelo de responsabilidade compartilhada da AWS
- Governance in the Nuvem AWS: The Right Balance Between Agility and Safety (Governança na Nuvem AWS: o equilíbrio certo entre agilidade e segurança)

OPS07-BP06 Viabilizar planos de suporte para workloads de produção

Viabilize o suporte para qualquer software e quaisquer serviços dos quais sua workload de produção dependa. Selecione um nível de suporte apropriado para atender às necessidades de nível de serviço da produção. Planos de suporte para essas dependências são necessários no caso de interrupção de um serviço ou de um problema de software. Documente os planos de suporte e como solicitar suporte para todos os fornecedores de serviços e software. Implemente mecanismos que verifiquem se os pontos de contato do suporte são mantidos atualizados.

Resultado desejado:

- Implemente planos de suporte para software e serviços dos quais as workloads de produção dependem.
- Escolha um plano de suporte apropriado com base nas necessidades de nível de serviço.
- Documente os planos de suporte, os níveis de suporte e como solicitar suporte.

Antipadrões comuns:

- Você não tem nenhum plano de suporte junto a um fornecedor de software essencial. Sua workload é afetada por isso e você não pode fazer nada para agilizar a correção ou obter atualizações em tempo hábil do fornecedor.
- Um desenvolvedor que era o principal ponto de contato com um fornecedor de software deixou a empresa. Você não consegue entrar em contato com o suporte do fornecedor diretamente. Você precisa despender tempo pesquisando e navegando por sistemas de contato genéricos, aumentando o tempo requerido para responder quando necessário.

- Ocorre uma interrupção na produção relacionada a um fornecedor de software. Não há nenhuma documentação sobre como abrir um caso de suporte.

Benefícios do estabelecimento desta prática recomendada:

- Com o nível de suporte apropriado, você é capaz de obter uma resposta no espaço de tempo requerido para atender a necessidades de nível de serviço.
- Como um cliente com suporte, você pode encaminhar a questão se houver problemas na produção.
- Os fornecedores de software e serviços podem ajudar na resolução de problemas durante um incidente.

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

Orientações para a implementação

Viabilize planos de suporte para qualquer software e quaisquer serviços dos quais sua workload de produção dependa. Estabeleça planos de suporte apropriados para atender a necessidades de nível de serviço. Para clientes da AWS, isso significa habilitar o AWS Business Support ou superior em quaisquer contas em que você tenha workloads de produção. Entre em contato regularmente com os fornecedores de suporte para obter atualizações sobre ofertas, processos e contatos de suporte. Documente como solicitar suporte de fornecedores de software e serviços e sobre como encaminhar problemas se houver uma interrupção. Implemente mecanismos para manter os contatos de suporte atualizados.

Exemplo de clientes

Na Loja UmaEmpresa, todas dependências de software e serviços comerciais contam com planos de suporte. Por exemplo, eles têm o AWS Enterprise Support habilitado em todas as contas com workloads de produção. Qualquer desenvolvedor pode abrir um caso de suporte quando há um problema. Há uma página de wiki com informações sobre como solicitar suporte, a quem notificar e as práticas recomendadas para agilizar um caso.

Etapas da implementação

1. Trabalhe com as partes interessadas em sua organização para identificar fornecedores de software e serviços dos quais sua workload dependa. Documente essas dependências.

2. Determine as necessidades de nível de serviço para sua workload. Selecione um plano de suporte alinhado a elas.
3. Para software e serviços comerciais, estabeleça um plano de suporte com os fornecedores.
    a. A assinatura do AWS Business Support ou superior para todas as contas de produção fornece um tempo de resposta rápido do AWS Support e é altamente recomendada. Se você não tiver suporte premium, precisará de um plano de ação para lidar com os problemas, o que requer a ajuda do AWS Support. O AWS Support oferece um conjunto de ferramentas e tecnologia, pessoas e programas projetados para ajudar você de forma proativa a otimizar a performance, reduzir custos e inovar com maior rapidez. O AWS Business Support oferece benefícios adicionais, incluindo acesso ao AWS Trusted Advisor e ao AWS Personal Health Dashboard e tempos de resposta mais rápidos.
4. Documente o plano de suporte em sua ferramenta de gerenciamentos de conhecimentos. Inclua como solicitar suporte, a quem notificar se for aberto um caso de suporte e como encaminhar o problema durante um incidente. Uma wiki é um bom mecanismo para possibilitar que todos façam as atualizações necessárias na documentação quando forem informados sobre alterações em processos ou contatos de suporte.

Nível de esforço do plano de implementação: baixo. A maioria dos fornecedores de software e serviços oferece planos de suporte que requerem adesão. Documentar e compartilhar práticas recomendadas no sistema de gerenciamento de conhecimentos garante que sua equipe saiba o que fazer quando houver um problema na produção.

Recursos

Práticas recomendadas relacionadas:

- OPS02-BP02 Processos e procedimentos com proprietários identificados

Documentos relacionados:

- AWS Support Plans

Serviços relacionados:

- AWS Business Support
- AWS Enterprise Support

Operar

Perguntas

- OPERAÇÕES 8. Como utilizar a observabilidade da workload em sua organização?
- OPERAÇÕES 9. Como compreender a integridade de suas operações?
- OPERAÇÕES 10. Como gerenciar os eventos de workload e operações?

OPERAÇÕES 8. Como utilizar a observabilidade da workload em sua organização?

Garanta a integridade ideal da workload usando a observabilidade. Utilize métricas, logs e rastreamentos relevantes para obter uma visão abrangente do desempenho de sua workload e resolver problemas com eficiência.

Práticas recomendadas

- OPS08-BP01 Analisar métricas de workload
- OPS08-BP02 Analisar logs de workloads
- OPS08-BP03 Analisar rastreamentos de workload
- OPS08-BP04 Criar alertas acionáveis
- OPS08-BP05 Criar painéis

OPS08-BP01 Analisar métricas de workload

Depois de implementar a telemetria de aplicações, analise regularmente as métricas coletadas. Embora a latência, as solicitações, os erros e a capacidade (ou cotas) forneçam informações sobre o desempenho do sistema, é fundamental priorizar a análise das métricas de resultados comerciais. Isso garante que você esteja tomando decisões orientadas por dados alinhadas aos seus objetivos de negócios.

Resultado desejado: Insights precisos sobre o desempenho da workload que impulsionam decisões baseadas em dados, garantindo o alinhamento com os objetivos de negócios.

Antipadrões comuns:

- Análise das métricas isoladamente, sem considerar seu impacto nos resultados comerciais.
- Confiança excessiva em métricas técnicas e, ao mesmo tempo, marginalização das métricas de negócios.

- Revisão pouco frequente das métricas, perdendo oportunidades de tomada de decisão em tempo real.

Benefícios de estabelecer esta prática recomendada:

- Compreensão aprimorada da correlação entre desempenho técnico e resultados comerciais.
- Processo de tomada de decisão aprimorado baseado em dados em tempo real.
- Identificação proativa e mitigação de problemas antes que eles afetem os resultados comerciais.

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

Orientação para implementação

Utilize ferramentas como o Amazon CloudWatch para realizar análises métricas. Serviços da AWS como o AWS Cost Anomaly Detection e o Amazon DevOps Guru podem ser usados para detectar anomalias, especialmente quando os limites estáticos são desconhecidos ou quando os padrões de comportamento são mais adequados para a detecção de anomalias.

Etapas da implementação

1. Analise e revise: Analise e interprete regularmente suas métricas de workload.
   a. Priorize as métricas de resultados comerciais em vez das métricas puramente técnicas.
   b. Entenda a importância de picos, quedas ou padrões em seus dados.
2. Use o Amazon CloudWatch: Use o Amazon CloudWatch para uma visão centralizada e uma análise aprofundada.
   a. Configure painéis do CloudWatch para visualizar suas métricas e compará-las ao longo do tempo.
   b. Use percentis no CloudWatch para obter uma visão clara da distribuição métrica, o que pode ajudar na definição de SLAs e na compreensão de valores discrepantes.
   c. Configure o AWS Cost Anomaly Detection para identificar padrões incomuns sem depender de limites estáticos.
   d. Implemente a observabilidade entre contas do CloudWatch para monitorar e solucionar problemas de aplicações que abrangem várias contas em uma região.
   e. Use insights métricos do CloudWatch para consultar e analisar dados métricos em contas e regiões, identificando tendências e anomalias.

   f. Aplique matemática métrica do CloudWatch para transformar, agregar ou realizar cálculos em suas métricas para obter insights mais profundos.

3. Empregue o Amazon DevOps Guru: Incorpore o Amazon DevOps Guru por sua detecção de anomalias aprimorada por machine learning para identificar sinais precoces de problemas operacionais em suas aplicações sem servidor e corrigi-los antes que afetem seus clientes.
4. Otimize com base em insights: Tome decisões informadas com base em sua análise métrica para ajustar e melhorar as workloads.

Nível de esforço do plano de implementação: Médio

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar os indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações

Documentos relacionados:

- The Wheel Blog - Emphasizing the importance of continually reviewing metrics (The Wheel Blog: como enfatizar a importância de revisar continuamente as métricas)
- Percentile are important (O percentil é importante)
- Using AWS Cost Anomaly Detection (Uso da AWS Cost Anomaly Detection)
- A observabilidade entre contas do CloudWatch
- Query your metrics with CloudWatch Metrics Insights (Consulte suas métricas com o CloudWatch Metrics Insights)

Vídeos relacionados:

- Enable Cross-Account Observability in Amazon CloudWatch (Ative a observabilidade entre contas no Amazon CloudWatch)
- Introduction to Amazon DevOps Guru (Introdução ao Amazon DevOps Guru)
- Continuously Analyze Metrics using AWS Cost Anomaly Detection (Analise continuamente as métricas usando o AWS Cost Anomaly Detection)

Exemplos relacionados:

- Um workshop de observabilidade
- Obter insights operacionais com AIOps usando Amazon DevOps Guru

OPS08-BP02 Analisar logs de workloads

Analisar regularmente os logs da workload é essencial para obter uma compreensão mais profunda dos aspectos operacionais de sua aplicação. Ao filtrar, visualizar e interpretar com eficiência os dados de log, você pode otimizar continuamente o desempenho e a segurança das aplicações.

Resultado desejado: Informações ricas sobre o comportamento e as operações da aplicação derivadas de uma análise completa de log, garantindo a detecção e mitigação proativas de problemas.

Antipadrões comuns:

- Negligenciar a análise dos logs até que surja um problema crítico.
- Não usar o conjunto completo de ferramentas disponíveis para análise de logs, perdendo insights essenciais.
- Confiar exclusivamente na revisão manual dos logs, sem aproveitar os recursos de automação e consulta.

Benefícios de estabelecer esta prática recomendada:

- Identificação proativa de gargalos operacionais, ameaças à segurança e outros possíveis problemas.
- Utilização eficiente dos dados de log para otimização contínua da aplicação.
- Compreensão aprimorada do comportamento da aplicação, auxiliando na depuração e solução de problemas.

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

Orientação para implementação

O Amazon CloudWatch Logs é uma ferramenta poderosa para análise de logs. Recursos integrados, como o CloudWatch Logs Insights e Contributor Insights, tornam intuitivo e eficiente o processo de derivação de informações significativas dos logs.

Etapas da implementação

1. Configure o CloudWatch Logs: Configure aplicações e serviços para enviar logs para o CloudWatch Logs.
2. Configure o CloudWatch Logs Insights: Use o CloudWatch Logs Insights para pesquisar e analisar interativamente seus dados de log.
   a. Crie consultas para extrair padrões, visualizar dados de log e obter insights acionáveis.
3. Utilize o Contributor Insights Use o CloudWatch Contributor Insights para identificar os principais locutores em dimensões de alta cardinalidade, como endereços IP ou agentes-usuários.
4. Implemente filtros de métrica do CloudWatch Logs: configure os filtros de métrica de log do CloudWatch para converter dados de log em métricas acionáveis. Isso permite que você defina alarmes ou analise melhor os padrões.
5. Revisão e refinamento regulares: Revise periodicamente suas estratégias de análise de log para capturar todas as informações relevantes e otimizar continuamente o desempenho da aplicação.

Nível de esforço do plano de implementação: Médio.

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar os indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações
- OPS08-BP01 Analisar métricas de workload

Documentos relacionados:

- Análise de dados de log com o CloudWatch Logs Insights
- Uso do CloudWatch Contributor Insights
- Criação e gerenciamento de filtros de métrica de log do CloudWatch Logs

Vídeos relacionados:

- Analyze Log Data with CloudWatch Logs Insights (Análise de dados de log com o CloudWatch Logs Insights)

- Use CloudWatch Contributor Insights to Analyze High-Cardinality Data (Use o CloudWatch Contributor Insights para analisar dados de alta cardinalidade)

Exemplos relacionados:

- Exemplos de consultas do CloudWatch Logs
- Um workshop de observabilidade

OPS08-BP03 Analisar rastreamentos de workload

Analisar dados de rastreamento é crucial para obter uma visão abrangente da jornada operacional de uma aplicação. Ao visualizar e compreender as interações entre vários componentes, o desempenho pode ser ajustado, os gargalos identificados e as experiências do usuário aprimoradas.

Resultado desejado: Obtenha uma visibilidade clara das operações distribuídas da sua aplicação, permitindo uma resolução mais rápida de problemas e uma experiência de usuário aprimorada.

Antipadrões comuns:

- Ignorar dados de rastreamento, confiando apenas em logs e métricas.
- Não correlacionar dados de rastreamento com logs associados.
- Ignorar as métricas derivadas de rastreamentos, como latência e taxas de falhas.

Benefícios de estabelecer esta prática recomendada:

- Aprimoramento da solução de problemas e redução do tempo médio de resolução (MTTR).
- Insights sobre dependências e seu impacto.
- Identificação e correção rápidas de problemas de desempenho.
- Uso de métricas derivadas de rastreamento para uma tomada de decisão informada.
- Experiências de usuário aprimoradas por meio de interações otimizadas de componentes.

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

Orientação para implementação

O AWS X-Ray oferece um pacote abrangente para análise de dados de rastreamento, fornecendo uma visão holística das interações de serviços, monitorando as atividades do usuário e detectando

problemas de desempenho. Recursos como ServiceLens, X-Ray Insights, X-Ray Analytics e Amazon DevOps Guru aprimoram a profundidade dos insights acionáveis derivados de dados de rastreamento.

Etapas da implementação

As etapas a seguir oferecem uma abordagem estruturada para implementar com eficácia a análise de dados de rastreamento usando serviços da AWS:

1. Integre o AWS X-Ray: Integre o X-Ray às suas aplicações para capturar dados de rastreamento.
2. Analise métricas do X-Ray: Aprofunde-se em métricas derivadas de rastreamentos do X-Ray, como latência, taxas de solicitação, taxas de falhas e distribuições de tempo de resposta usando o mapa de serviços para monitorar a integridade da aplicação.
3. Use o ServiceLens: Use o mapa do ServiceLens para melhorar a observabilidade de seus serviços e aplicações. Isso permite a visualização integrada de rastreamentos, métricas, logs, alarmes e outras informações de integridade.
4. Habilite o X-Ray Insights:
   a. Ative o X-Ray Insights para detecção automática de anomalias em rastreamentos.
   b. Examine os insights para identificar padrões e determinar as causas principais, como maiores taxas de falhas ou latências.
   c. Consulte o cronograma de insights para uma análise cronológica dos problemas detectados.
5. Use o X-Ray Analytics: O X-Ray Analytics permite que você explore minuciosamente os dados de rastreamento, identifique padrões e extraia insights.
6. Use grupos no X-Ray: Crie grupos no X-Ray para filtrar rastreamentos com base em critérios como alta latência, permitindo uma análise mais direcionada.
7. Incorpore o Amazon DevOps Guru: Use o Amazon DevOps Guru para se beneficiar dos modelos de machine learning que identificam anomalias operacionais nos rastreamentos.
8. Use o CloudWatch Synthetics: Use o CloudWatch Synthetics para criar canários para monitorar continuamente os endpoints e fluxos de trabalho. Esses canários podem integrar-se com o X-Ray para fornecer dados de rastreamento para uma análise aprofundada das aplicações que estão sendo testadas.
9. Use o Monitoramento de Usuários Reais (RUM): Com o AWS X-Ray e o CloudWatch RUM, você pode analisar e depurar o caminho da solicitação a partir dos usuários finais de sua aplicação por meio de serviços downstream gerenciados pela AWS . Isso ajuda você a identificar tendências e erros de latência que afetam seus usuários.

10. Correlacionar com logs: Correlacione dados de rastreamento com logs relacionados dentro da visualização de rastreamento do X-Ray para uma perspectiva granular sobre o comportamento da aplicação. Isso permite que você visualize eventos de log diretamente associados às transações rastreadas.

Nível de esforço do plano de implementação: Médio.

Recursos

Práticas recomendadas relacionadas:

- OPS08-BP01 Analisar métricas de workload
- OPS08-BP02 Analisar logs de workloads

Documentos relacionados:

- Uso do ServiceLens para monitorar a integridade da aplicação
- Explorar dados de rastreamento com o X-Ray Analytics
- Detectar anomalias em rastreamentos com o X-Ray Insights
- Monitorar continuamente com o CloudWatch Synthetics

Vídeos relacionados:

- Analyze and Debug Applications Using Amazon CloudWatch Synthetics and AWS X-Ray (Analise e depure aplicações usando Amazon CloudWatch Synthetics e AWS X-Ray)
- Use AWS X-Ray Insights (Use o AWS X-Ray Insights)

Exemplos relacionados:

- Um workshop de observabilidade
- Como implementar o X-Ray com o AWS Lambda
- Modelos canário do CloudWatch Synthetics

OPS08-BP04 Criar alertas acionáveis

Detectar e responder prontamente aos desvios no comportamento da sua aplicação é crucial. É essencial reconhecer quando os resultados com base nos indicadores-chave de performance (KPIs) estão em risco ou quando surgem anomalias inesperadas. Basear alertas em KPIs garante que os sinais que você recebe estejam diretamente vinculados ao impacto comercial ou operacional. Essa abordagem de alertas acionáveis promove respostas proativas e ajuda a manter o desempenho e a confiabilidade do sistema.

Resultado desejado: Receba alertas oportunos, relevantes e acionáveis para rápida identificação e mitigação de possíveis problemas, especialmente quando os resultados do KPI estão em risco.

Antipadrões comuns:

- A configuração de muitos alertas não críticos leva à fadiga de alertas.
- A não priorização de alertas com base em KPIs dificulta a compreensão do impacto comercial dos problemas.
- A não abordagem das causas-raiz leva a alertas repetitivos para o mesmo problema.

Benefícios de estabelecer esta prática recomendada:

- Redução da fadiga de alertas ao se concentrar em alertas acionáveis e relevantes.
- Maior disponibilidade e confiabilidade do sistema por meio da detecção e mitigação proativas de problemas.
- Colaboração em equipe aprimorada e resolução mais rápida de problemas por meio da integração com ferramentas populares de alerta e comunicação.

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

Orientação para implementação

Para criar um mecanismo de alerta eficaz, é fundamental usar métricas, logs e dados de rastreamento que sinalizem quando os resultados com base nos KPIs estão em risco ou quando anomalias são detectadas.

Etapas da implementação

1. Determine indicadores-chave de performance (KPIs): Identifique os KPIs de sua aplicação. Os alertas devem estar vinculados a esses KPIs para refletir com precisão o impacto nos negócios.

2. Implemente a detecção de anomalias:
    - Use o AWS Cost Anomaly Detection: configure o AWS Cost Anomaly Detection para detectar automaticamente padrões incomuns, garantindo que os alertas sejam gerados somente para anomalias genuínas.
    - Use o X-Ray Insights:
        a. Configure o X-Ray Insights para detectar anomalias nos dados de rastreamento.
        b. Configure notificações no X-Ray Insights para ser alertado sobre problemas detectados.
    - Integre com o DevOps Guru:
        a. Utilize o Amazon DevOps Guru devido a seus recursos de machine learning na detecção de anomalias operacionais com dados existentes.
        b. Navegue até as configurações de notificação no DevOps Guru para configurar alertas de anomalias.
3. Implemente alertas acionáveis: Crie alertas que forneçam informações adequadas para ação imediata.
4. Reduza a fadiga de alarmes: Minimize os alertas não críticos. Equipes sobrecarregadas com vários alertas insignificantes podem não perceber problemas críticos e a eficácia geral do mecanismo de alerta fica diminuída.
5. Configurar alarmes compostos: Use os alarmes compostos do Amazon CloudWatch para consolidar vários alarmes.
6. Integre com ferramentas de alerta: Incorpore ferramentas como Ops Genie e PagerDuty.
7. Utilize o AWS Chatbot integre o AWS Chatbotpara retransmitir alertas para Chime, Microsoft Teams e Slack.
8. Alerta baseado em logs: Use o filtros de métrica de log no CloudWatch para criar alarmes com base em eventos de log específicos.
9. Revise e repita: Revise e revisite regularmente as configurações de alerta.

Nível de esforço do plano de implementação: Médio.

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar os indicadores-chave de performance
- OPS04-BP02 Implementar a telemetria de aplicações

- OPS04-BP03 Implementar a telemetria da experiência do usuário
- OPS04-BP04 Implementar a telemetria de dependências
- OPS04-BP05 Implementar rastreamento distribuído
- OPS08-BP01 Analisar métricas de workload
- OPS08-BP02 Analisar logs de workloads
- OPS08-BP03 Analisar rastreamentos de workload

Documentos relacionados:

- Uso dos alarmes do Amazon CloudWatch
- Crie um alarme composto
- Crie um alarme do CloudWatch com base na detecção de anomalias
- Notificações do DevOps Guru
- Notificações do X-Ray Insights
- Monitore, opere e solucione problemas de seus recursos da AWS com ChatOps interativos
- Guia de integração do Amazon CloudWatch | PagerDuty
- Integre o OpsGenie com o Amazon CloudWatch

Vídeos relacionados:

- Create Composite Alarms in Amazon CloudWatch (Criar alarmes compostos no Amazon CloudWatch)
- AWS Chatbot Overview (Visão geral do AWS Chatbot)
- AWS on Air ft. Mutative Commands in AWS Chatbot (Comandos mutativos no AWS Chatbot)

Exemplos relacionados:

- Alarmes, gerenciamento de incidentes e remediação na nuvem com o Amazon CloudWatch
- Tutorial: criação de uma regra do Amazon EventBridge que envia notificações para o AWS Chatbot
- Um workshop de observabilidade

OPS08-BP05 Criar painéis

Os painéis são a visão centrada no ser humano dos dados de telemetria de suas workloads. Embora forneçam uma interface visual vital, eles não devem substituir os mecanismos de alerta, mas sim complementá-los. Quando elaborados com cuidado, eles não apenas oferecem insights rápidos sobre a integridade e o desempenho do sistema, como também podem apresentar às partes interessadas informações em tempo real sobre os resultados empresariais e o impacto dos problemas.

Resultado desejado: Insights claros e acionáveis sobre a integridade do sistema e dos negócios usando representações visuais.

Antipadrões comuns:

- Painéis complicados demais e com muitas métricas.
- Confiar em painéis sem alertas para detecção de anomalias.
- Não atualizar os painéis à medida que as workloads evoluem.

Benefícios de estabelecer esta prática recomendada:

- Visibilidade imediata das métricas e KPIs críticos do sistema.
- Comunicação e compreensão aprimoradas com as partes interessadas.
- Visão rápida do impacto dos problemas operacionais.

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

Orientação para implementação

Painéis centrados nos negócios

Painéis personalizados para os KPIs de negócios envolvem uma gama maior de partes interessadas. Embora essas pessoas possam não estar interessadas nas métricas do sistema, elas estão interessadas em entender as implicações comerciais desses números. Um painel centrado nos negócios garante que todas as métricas técnicas e operacionais monitoradas e analisadas estejam sincronizadas com as metas empresariais abrangentes. Esse alinhamento fornece clareza, garantindo que todos estejam em sintonia sobre o que é essencial e o que não é. Além disso, painéis que destacam os KPIs de negócios tendem a ser mais acionáveis. As partes interessadas podem

entender rapidamente a integridade das operações, as áreas que precisam de atenção e o impacto potencial nos resultados empresariais.

Com isso em mente, ao criar seus painéis, garanta que haja um equilíbrio entre métricas técnicas e KPIs comerciais. Ambos são vitais, mas atendem a públicos diferentes. O ideal é que você tenha painéis que forneçam uma visão holística da integridade e do desempenho do sistema e, ao mesmo tempo, enfatizem os principais resultados comerciais e suas implicações.

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 da AWS e contas diferentes.

Etapas da implementação

1. Crie um painel básico: crie um novo painel no CloudWatche dê a ele um nome descritivo.
2. Use widgets de Markdown: antes de mergulhar nas métricas, use widgets de Markdown para adicionar contexto textual na parte superior do painel. Isso deve explicar o que o painel abrange, a importância das métricas representadas e também pode conter links para outros painéis e ferramentas de solução de problemas.
3. Crie variáveis do painel: incorpore variáveis do painel quando apropriado, para permitir visualizações dinâmicas e flexíveis do painel.
4. Crie widgets de métricas: adicione widgets de métricas para visualizar várias métricas que sua aplicação emite, adaptando esses widgets para representar com eficácia a integridade do sistema e os resultados empresariais.
5. Consultas do Log Insights: utilize o CloudWatch Logs Insights para obter métricas acionáveis de seus logs e exibir esses insights em seu painel.
6. Configurar alarmes: integre alarmes do CloudWatch em seu painel para uma visão rápida de qualquer métrica que esteja ultrapassando seus limites.
7. Use o Contributor Insights: incorpore o CloudWatch Contributor Insights para analisar campos de alta cardinalidade e obter uma compreensão mais clara dos principais colaboradores do seu recurso.
8. Crie widgets personalizados: para necessidades específicas não atendidas pelos widgets padrão, considere criar widgets personalizados. Eles podem ser extraídos de várias fontes de dados ou representar dados de maneiras exclusivas.
9. Repita e refine: à medida que sua aplicação evolui, revise regularmente seu painel para garantir sua relevância.

Recursos

Práticas recomendadas relacionadas:

- OPS04-BP01 Identificar os indicadores-chave de performance
- OPS08-BP01 Analisar métricas de workload
- OPS08-BP02 Analisar logs de workloads
- OPS08-BP03 Analisar rastreamentos de workload
- OPS08-BP04 Criar alertas acionáveis

Documentos relacionados:

- Criação de painéis para visibilidade operacional
- Uso de painéis do Amazon CloudWatch

Vídeos relacionados:

- Create Cross Account & Cross Region CloudWatch Dashboards (Criar painéis do CloudWatch entre contas e entre regiões)
- AWS re:Invent 2021 - Gain enterprise visibility with Nuvem AWS operation dashboards (AWS re:Invent 2021: obtenha visibilidade corporativa com painéis de operação do CloudWatch)

Exemplos relacionados:

- Um workshop de observabilidade
- Monitoramento de aplicações do Amazon CloudWatch

OPERAÇÕES 9. Como compreender 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 Medir metas operacionais e KPIs com métricas
- OPS09-BP02 Comunicar o status e as tendências para garantir a visibilidade da operação
- OPS09-BP03 Revisar as métricas operacionais e priorizar a melhoria

OPS09-BP01 Medir metas operacionais e KPIs com métricas

Obtenha metas e KPIs que definam o sucesso das operações de sua organização e determine se as métricas os refletem. Defina linhas de base como ponto de referência e reavalie regularmente. Desenvolva mecanismos para coletar essas métricas das equipes para avaliação.

Resultado desejado:

- As metas e os KPIs das equipes de operações da organização foram publicados e compartilhados.
- Métricas que refletem esses KPIs são estabelecidas. Os exemplos podem incluir:
  - Profundidade da fila de tíquetes ou idade média do tíquete
  - Contagem de tíquetes agrupada por tipo de problema
  - Tempo gasto trabalhando em problemas com ou sem um procedimento operacional padronizado (SOP)
  - Tempo gasto na recuperação de uma falha no envio de código
  - Volume de chamadas

Antipadrões comuns:

- Os prazos de implantação são perdidos porque os desenvolvedores são contratados para realizar tarefas de solução de problemas. As equipes de desenvolvimento demandam mais pessoal, mas não conseguem quantificar quantos precisam porque o tempo perdido não pode ser medido.
- Um atendimento de Nível 1 foi configurado para lidar com chamadas de usuários. Com o tempo, mais workloads foram adicionadas, mas nenhum número de funcionários foi alocado para o atendimento de Nível 1. A satisfação do cliente sofre à medida que os tempos de atendimento aumentam e os problemas ficam mais tempo sem resolução, mas a gerência não vê indicadores disso, impedindo qualquer ação.
- Uma workload problemática foi transferida para uma equipe de operações separada para manutenção. Diferentemente de outras workloads, a nova não foi fornecida com documentação e runbooks adequados. Dessa forma, as equipes passam mais tempo solucionando problemas e tratando de falhas. No entanto, não há métricas que documentem isso, o que dificulta a prestação de contas.

Benefícios de estabelecer esta prática recomendada: Onde o monitoramento da workload mostra o estado de nossas aplicações e serviços, as equipes de operações de monitoramento fornecem aos proprietários uma visão das mudanças entre os consumidores dessas workloads, como as

mudanças nas necessidades dos negócios. Meça a eficácia dessas equipes e avalie-as em relação às metas de negócios, criando métricas que possam refletir o estado das operações. As métricas podem destacar problemas de suporte ou identificar quando ocorrem desvios de uma meta de nível de serviço.

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

Orientação para implementação

Agende um horário com líderes de negócios e partes interessadas para determinar as metas gerais do serviço. Determine quais devem ser as tarefas de várias equipes de operações e quais desafios elas podem enfrentar. Com isso, pense em indicadores-chave de performance (KPIs) que possam refletir essas metas operacionais. Pode ser a satisfação do cliente, o tempo desde a concepção do recurso até a implantação, o tempo médio de resolução de problemas e outros.

Trabalhando a partir de KPIs, identifique as métricas e as fontes de dados que podem refletir melhor essas metas. A satisfação do cliente pode ser uma combinação de várias métricas, como tempos de espera ou resposta de chamadas, índices de satisfação e tipos de problemas levantados. Os tempos de implantação podem ser a soma do tempo necessário para testes e implantação e quaisquer correções pós-implantação que precisem ser adicionadas. As estatísticas que mostram o tempo gasto em diferentes tipos de problemas (ou a contagem desses problemas) podem fornecer uma visão de onde é necessário um esforço direcionado.

Recursos

Documentos relacionados:

- Amazon QuickSight: uso de KPIs
- Amazon CloudWatch: uso de métricas
- Criação de painéis
- Como rastrear seus KPIs de otimização de custos com o painel de KPI

OPS09-BP02 Comunicar o status e as tendências para garantir a visibilidade da operação

É necessário conhecer o estado de suas operações e a direção das tendências para identificar quando os resultados podem estar em risco, se o trabalho adicional pode ou não ser apoiado ou os efeitos que as mudanças tiveram em suas equipes. Durante eventos operacionais, ter páginas de status que os usuários e as equipes operacionais possam consultar para obter informações pode reduzir a pressão nos canais de comunicação e disseminar informações de forma proativa.

Resultado desejado:

- Os líderes de operações têm uma visão rápida para ver em que tipo de volume de chamadas suas equipes estão operando e quais esforços podem estar em andamento, como implantações.
- Os alertas são disseminados para as partes interessadas e comunidades de usuários quando ocorrem impactos nas operações normais.
- A liderança da organização e as partes interessadas podem verificar uma página de status em resposta a um alerta ou impacto e obter informações sobre um evento operacional, como pontos de contato, informações sobre tíquetes e tempos estimados de recuperação.
- Os relatórios são disponibilizados para a liderança e outras partes interessadas para mostrar estatísticas operacionais, como volumes de chamadas durante um período de tempo, índices de satisfação do usuário, números de tíquetes pendentes e suas idades.

Antipadrões comuns:

- Uma workload diminui, deixando um serviço indisponível. O volume de chamadas aumenta à medida que os usuários solicitam saber o que está acontecendo. Os gerentes aumentam o volume de solicitações para saber quem está resolvendo um problema. Várias equipes de operações duplicam esforços na tentativa de investigar.
- O desejo por uma nova capacidade faz com que vários funcionários sejam transferidos para um esforço de engenharia. Nenhum preenchimento é fornecido e os tempos de resolução de problemas aumentam. Essas informações não são capturadas e a liderança toma conhecimento do problema somente após várias semanas de comentários de insatisfação do usuário.

Benefícios de estabelecer esta prática recomendada: durante eventos operacionais em que a empresa é afetada, muito tempo e energia podem ser desperdiçados consultando informações de várias equipes tentando entender a situação. Ao estabelecer páginas de status e painéis amplamente divulgados, as partes interessadas podem obter rapidamente informações, como se um problema foi detectado ou não, quem liderou o problema ou quando é esperado um retorno às operações normais. Isso permite que os membros da equipe dediquem mais tempo à resolução de problemas e passem menos tempo comunicando o status a outras pessoas.

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

Orientação para implementação

Crie painéis que mostrem as principais métricas atuais para suas equipes de operações e torne-as facilmente acessíveis, tanto para os líderes de operações quanto para a gerência.

Crie páginas de status que possam ser atualizadas rapidamente para mostrar quando um incidente ou evento está ocorrendo, quem é o proprietário e quem está coordenando a resposta. Compartilhe todas as etapas ou soluções alternativas que os usuários devem considerar nesta página e divulgue amplamente a localização. Incentive os usuários a verificar esse local primeiro quando confrontados com um problema desconhecido.

Colete e forneça relatórios que mostrem a integridade das operações ao longo do tempo e distribua-os aos líderes e tomadores de decisão para ilustrar o trabalho das operações junto com os desafios e as necessidades.

Compartilhe entre as equipes essas métricas e relatórios que melhor refletem as metas e os KPIs e onde eles foram influentes na promoção da mudança. Dedique tempo a essas atividades para aumentar a importância das operações dentro das equipes e entre elas.

Recursos

Documentos relacionados:

- Avalie o progresso
- Criação de painéis para visibilidade da operação

Soluções relacionadas:

- Operações de dados

OPS09-BP03 Revisar as métricas operacionais e priorizar a melhoria

Reservar tempo e dedicar recursos para analisar o estado das operações garante que atender à linha de negócios do dia a dia continue sendo uma prioridade. Reúna líderes de operações e partes interessadas para revisar regularmente as métricas, reafirmar ou modificar metas e objetivos e priorizar melhorias.

Resultado desejado:

- Os líderes de operações e a equipe se reúnem regularmente para revisar as métricas durante um determinado período do relatório. Os desafios são comunicados, as vitórias são celebradas e as lições aprendidas são compartilhadas.
- As partes interessadas e os líderes de negócios são regularmente informados sobre o estado das operações e solicitados a fornecer informações sobre metas, KPIs e iniciativas futuras. As compensações entre prestação de serviços, operações e manutenção são discutidas e contextualizadas.

Antipadrões comuns:

- Um novo produto é lançado, mas as equipes operacionais de nível 1 e nível 2 não são adequadamente treinadas para dar suporte nem recebem pessoal adicional. Métricas que mostram a diminuição nos tempos de resolução de tíquetes e o aumento nos volumes de incidentes não são vistas pelos líderes. Uma ação é tomada semanas depois, quando os números de assinaturas começam a cair à medida que usuários insatisfeitos saem da plataforma.
- Um processo manual para realizar a manutenção de uma workload está em vigor há muito tempo. Embora o desejo de automatizar estivesse presente, essa era uma prioridade baixa, dada a baixa importância do sistema. No entanto, com o tempo, o sistema cresceu em importância e agora esses processos manuais consomem a maior parte do tempo das operações. Nenhum recurso está programado para fornecer mais ferramentas às operações, causando o esgotamento da equipe à medida que as workloads aumentam. A liderança percebe o que está acontecendo quando é relatado que funcionários estão indo trabalhar para outros concorrentes.

Benefícios de estabelecer esta prática recomendada: em algumas organizações, pode ser um desafio alocar o mesmo tempo e atenção dedicados à prestação de serviços e a novos produtos ou ofertas. Quando isso ocorre, a linha de negócios pode sofrer enquanto o nível de serviço esperado se deteriora lentamente. Isso ocorre porque as operações não mudam e evoluem com o crescimento dos negócios e logo podem ser deixadas para trás. Sem uma análise regular dos insights que as operações coletam, o risco para a empresa pode se tornar visível somente quando for tarde demais. Ao alocar tempo para revisar métricas e procedimentos tanto entre a equipe de operações quanto com a liderança, o papel crucial que as operações desempenham permanece visível e os riscos podem ser identificados muito antes de atingirem níveis críticos. As equipes de operações obtêm uma visão melhor das mudanças e iniciativas comerciais iminentes, permitindo que esforços proativos sejam realizados. A visibilidade da liderança nas métricas operacionais mostra o papel que essas equipes desempenham na satisfação do cliente, tanto interna quanto externa, e permite

que elas avaliem melhor as opções de prioridades ou garantam que as operações tenham tempo e recursos para mudar e evoluir com novas iniciativas de negócios e workload.

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

Orientação para implementação

Dedique tempo para analisar as métricas operacionais entre as partes interessadas e as equipes operacionais e analisar os dados do relatório. Coloque esses relatórios nos contextos das metas e objetivos da organização para determinar se eles estão sendo cumpridos. Identifique fontes de ambiguidade onde as metas não são claras ou onde pode haver conflitos entre o que é pedido e o que é dado.

Identifique onde o tempo, as pessoas e as ferramentas podem ajudar nos resultados das operações. Determine quais KPIs isso afetaria e quais deveriam ser as metas de sucesso. Revise regularmente para garantir que as operações tenham recursos suficientes para apoiar a linha de negócios.

Recursos

Documentos relacionados:

- Amazon Athena
- Amazon CloudWatch metrics and dimensions reference (Métricas do Amazon CloudWatch e referência de dimensões)
- Amazon QuickSight
- AWS Glue
- AWS Glue Data Catalog
- Collect metrics and logs from Amazon EC2 instances and on-premises servers with the Amazon CloudWatch Agent (Coletar métricas e logs das instâncias do Amazon EC2 e de servidores on-premises com o agente do Amazon CloudWatch)
- Using Amazon CloudWatch metrics (Uso de métricas do Amazon CloudWatch)

OPERAÇÕES 10. Como gerenciar os eventos de workload 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

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

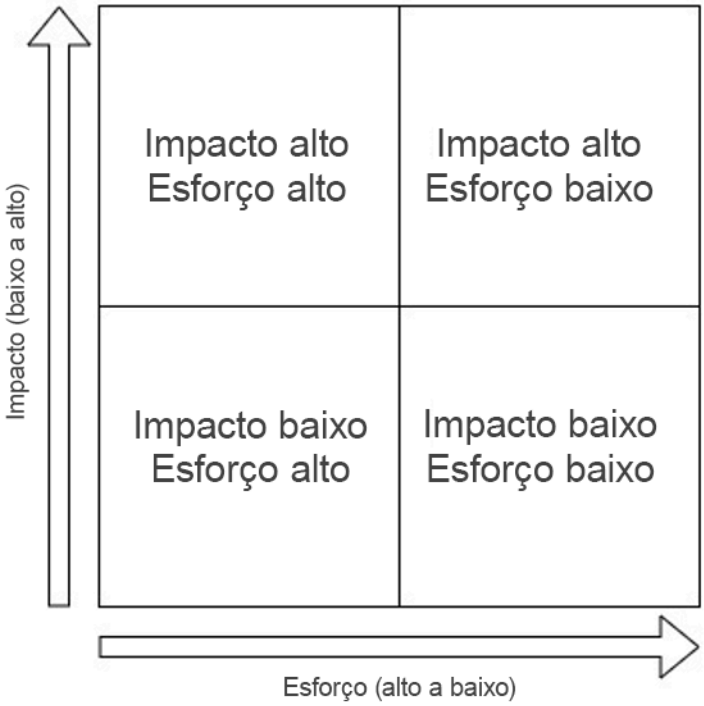
A AnyCompany Retail tem uma parte de sua wiki interna dedicada a processos para gerenciamento de eventos, incidentes e problemas. Todos os eventos são enviados para o Amazon EventBridge. Os problemas são identificados como OpsItems no OpsCenter do AWS Systems Manager e priorizados para correção, reduzindo 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 Definir um plano de comunicação com o cliente para interrupções

Defina e teste um plano de comunicação para interrupções do sistema que seja confiável para manter os clientes e as partes interessadas informados durante interrupções. Comunique-se diretamente com os usuários tanto quando os serviços que eles usam forem afetados como quando os serviços voltarem ao normal.

Resultado desejado:

- Você tem um plano de comunicação para situações que vão desde manutenção agendada até grandes falhas inesperadas, incluindo invocação de planos de recuperação de desastres.
- Nas comunicações, você fornece informações claras e transparentes sobre problemas do sistema para ajudar os clientes a evitar dúvidas sobre o desempenho dos sistemas.
- Você usa mensagens de erro personalizadas e páginas de status para reduzir o pico nas solicitações de suporte técnico e mantém os usuários informados.
- O plano de comunicação é testado regularmente para verificar se ele ocorrerá como planejado no caso de uma interrupção real.

Antipadrões comuns:

- Ocorre uma interrupção da workload, mas você não tem um plano de comunicação. Os usuários sobrecarregam o sistema de tíquetes com solicitações, pois não têm informações sobre a interrupção.
- Você envia uma notificação por e-mail aos usuários durante uma interrupção. Ela não contém um prazo para a restauração do serviço, então os usuários não conseguem se planejar em torno da interrupção.
- Há um plano de comunicação para interrupções, mas ele nunca foi testado. Ocorre uma interrupção e o plano de comunicação falha, pois faltou uma etapa fundamental que poderia ter sido identificada no teste.
- Durante uma interrupção, você envia uma notificação aos usuários com muitas informações e detalhes técnicos sob o NDA da AWS.

Benefícios do estabelecimento desta prática recomendada:

- Manter a comunicação durante as interrupções garante que os clientes possam ver o andamento da resolução dos problemas e o tempo previsto para que ela ocorra.

- Desenvolver um plano de comunicação bem-definido garante que os clientes e usuários finais estejam bem-informados para que possam tomar medidas adicionais visando a mitigar o impacto das interrupções.
- Com uma comunicação adequada e maior ciência acerca de interrupções planejadas e não planejadas, é possível melhorar a satisfação dos clientes, limitar reações não pretendidas e gerar a retenção dos clientes.
- Uma comunicação transparente e em tempo hábil acerca da interrupção do sistema gera credibilidade e estabelece a confiança necessária para manter seu relacionamento com os clientes.
- Uma estratégia de comunicação comprovada durante uma interrupção ou crise reduz a especulação e os rumores que poderiam atrapalhar sua capacidade de recuperação.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientações para a implementação

Os planos de comunicação que mantêm os clientes informados durante interrupções são holísticos e abrangem várias interfaces, incluindo páginas de erro voltadas para o cliente, mensagens de erro de API personalizadas, banners sobre o status do sistema e páginas de status de integridade. Se o sistema incluir usuários registrados, é possível comunicar-se por canais de mensagens, como e-mail, SMS ou notificações por push, para enviar conteúdo com mensagens personalizadas aos clientes.

Ferramentas de comunicação com o cliente

Como uma primeira linha de defesa, as aplicações web e móveis devem fornecer mensagens de erro amistosas e informativas durante uma interrupção e devem poder redirecionar o tráfego para uma página de status. O Amazon CloudFront é uma rede de entrega de conteúdo (CDN) que inclui recursos para definir e entregar conteúdo de erro personalizado. As páginas de erro personalizadas no CloudFront são uma ótima camada inicial de mensagens para os clientes para interrupções no nível de componentes. O CloudFront também pode simplificar o gerenciamento e a ativação da página de status para interceptar todas as solicitações durante interrupções planejadas e não planejadas.

As mensagens de erro de API personalizadas podem ajudar a detectar e reduzir o impacto quando as interrupções são isoladas a serviços discretos. O Amazon API Gateway permite configurar respostas personalizadas para as APIs REST. Isso permite fornecer mensagens claras e significativas para os consumidores da API quando o API Gateway não puder acessar os serviços de back-end. As mensagens personalizadas também podem ser usadas para dar suporte a notificações

e conteúdos de banner sobre a interrupção quando um recurso específico do sistema é danificado devido a interrupções no nível do serviço.

As mensagens diretas são o tipo mais personalizado de mensagens para o cliente. O Amazon Pinpoint é um serviço gerenciado para comunicações escaláveis de vários canais. O Amazon Pinpoint permite criar campanhas que possam transmitir mensagens amplamente pela base de clientes afetados por SMS, e-mail, mensagem de voz, notificações por push ou canais personalizados definidos por você. Ao gerenciar as mensagens com o Amazon Pinpoint, as campanhas de mensagem são bem-definidas, testáveis e podem ser aplicadas de forma inteligente a segmentos de clientes-alvo. Depois de serem estabelecidas, as campanhas podem ser agendadas ou acionadas por eventos e podem ser facilmente testadas.

Exemplo de clientes

Quando a workload é prejudicada, a Loja UmaEmpresa envia uma notificação por e-mail aos usuários. O e-mail descreve qual funcionalidade da empresa foi prejudicada e fornece uma estimativa realista de quando o serviço será restaurado. Além disso, há uma página de status que mostra informações em tempo real sobre a integridade da workload. O plano de comunicação é testado em um ambiente de desenvolvimento duas vezes ao ano para validar sua eficácia.

Etapas da implementação

1. Determine os canais de comunicação para sua estratégia de mensagens. Considere os aspectos da arquitetura da aplicação e determine a melhor estratégia para fornecer feedback aos clientes. Isso pode incluir uma ou mais das estratégias de orientação descritas, incluindo páginas de erro e de status, respostas de erro de API personalizadas ou mensagens diretas.
2. Elabore páginas de status para a aplicação. Se você determinou que as páginas de status ou de erro personalizadas são adequadas para os clientes, é necessário elaborar o conteúdo e as mensagens para essas páginas. As páginas de erro explicam aos usuários por que uma aplicação não está disponível, quando ela pode ficar disponível novamente e o que pode ser feito enquanto isso. Se a aplicação usar o Amazon CloudFront, é possível fornecer respostas de erro personalizadas ou usar o Lambda no Edge para traduzir erros e reescrever o conteúdo da página. O CloudFront também permite mudar os destinos do conteúdo da aplicação para uma origem de conteúdo estático do Amazon S3 que contém sua página de status da interrupção ou de manutenção.
3. Elabore o conjunto de status de erro de API correto para seu serviço. As mensagens de erro produzidas pelo API Gateway quando ele não consegue acessar os serviços de back-end, além das exceções no nível do serviço, podem não conter mensagens amistosas adequadas para

exibição aos usuários finais. Sem precisar fazer alterações no código dos serviços de back-end, é possível configurar as respostas de erro personalizadas do API Gateway para mapear os códigos de resposta HTTP para mensagens de erro de API selecionadas.

4. Elabore mensagens de uma perspectiva empresarial para que elas sejam relevantes aos usuários finais do sistema e não contenham detalhes técnicos. Considere seu público e alinhe suas mensagens. Por exemplo, você pode conduzir os usuários internos para uma solução alternativa ou um processo manual que utiliza sistemas alternativos. Os usuários externos podem ser solicitados a aguardar até que o sistema seja restaurado ou assinar as atualizações para receber uma notificação quando o sistema for restaurado. Defina mensagens aprovadas para vários cenários, incluindo interrupções não planejadas, manutenção planejada e falhas parciais do sistema quando um recurso específico pode estar danificado ou indisponível.
5. Modele e automatize as mensagens para os clientes. Depois de estabelecer o conteúdo das mensagens, é possível usar o Amazon Pinpoint ou outras ferramentas para automatizar sua campanha de mensagens. Com o Amazon Pinpoint, é possível criar segmentos de destino de clientes para usuários afetados específicos e transformar as mensagens em modelos. Consulte o Tutorial do Amazon Pinpoint para entender como configurar uma campanha de mensagens.
6. Evite o acoplamento forte de recursos de mensagens ao sistema voltado para o cliente. Sua estratégia de mensagens não deve depender fortemente de serviços ou armazenamentos de dados do sistema para verificar se é possível enviar mensagens quando ocorrerem interrupções. Considere desenvolver a capacidade de enviar mensagens a mais de uma região ou zona de disponibilidade para disponibilidade de mensagens. Se você estiver usando os serviços da AWS para enviar mensagens, utilize as operações do plano de dados sobre as operações do ambiente de gerenciamento para invocar suas mensagens.

Nível de esforço do plano de implementação: alto. Desenvolver um plano de comunicação e os mecanismos para enviá-lo pode exigir um esforço significativo.

Recursos

Práticas recomendadas relacionadas:

- OPS07-BP03 Usar runbooks para realizar procedimentos – Seu plano de comunicação deve ter um runbook associado a ele para que seus funcionários saibam como responder.
- OPS11-BP02 Executar análise pós-incidente – Depois de uma interrupção, realize uma análise pós-incidente para identificar mecanismos, a fim de evitar outra interrupção.

Documentos relacionados:

- Error Handling Patterns in Amazon API Gateway and AWS Lambda(Padrões de tratamento de erros no Amazon API Gateway e no AWS Lambda)
- Amazon API Gateway responses (Respostas do Amazon API Gateway)

Exemplos relacionados:

- AWS Health Dashboard (Painel do AWS Health)
- Summary of the AWS Service Event in the Northern Virginia (US-EAST-1) Region (Resumo do evento de serviço da AWS na região Virgínia do Norte (US-EAST-1)

Serviços relacionados:

- AWS Support
- Contrato de Cliente da AWS
- Amazon CloudFront
- Amazon API Gateway
- Amazon Pinpoint
- Amazon S3

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

- OPERAÇÕES 11. Como evoluir as operações?

OPERAÇÕES 11. Como evoluir as operações?

Dedique tempo e recursos para a melhoria incremental praticamente 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 Realizar 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 sua workload em relação às práticas recomendadas de arquitetura interna e externa. Realize análises da workload pelo menos uma vez ao ano. Priorize as oportunidades de melhoria em sua cadência de desenvolvimento de software.

Resultado desejado:

- Você analisa sua workload em relação às práticas recomendadas de arquitetura pelo menos anualmente.
- As oportunidades de melhoria recebem a mesma prioridade em seu processo de desenvolvimento de software.

Antipadrões comuns:

- Você não realizou uma análise de arquitetura em sua workload desde que foi implantada há vários anos.
- As oportunidades de melhoria recebem uma prioridade mais baixa e permanecem no backlog.
- Não há um padrão para implementar modificações nas práticas recomendadas da organização.

Benefícios do estabelecimento desta prática recomendada:

- Sua workload é mantida atualizada em relação às práticas recomendadas de arquitetura.
- A evolução de sua workload é realizada de forma deliberada.
- Você pode utilizar as práticas recomendadas da organização para melhorar todas as workloads.

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

Orientações para a implementação

Pelo menos anualmente, você realiza uma análise arquitetônica de sua workload. Usando práticas recomendadas internas e externas, avalie sua workload e identifique oportunidades de melhoria. Priorize as oportunidades de melhoria em sua cadência de desenvolvimento de software.

Exemplo de clientes

Todas as workloads da AnyCompany Retail passam por um processo anual de análise da arquitetura. A empresa desenvolveu a própria lista de verificação de práticas recomendadas que se aplicam a todas as workloads. Usando o recurso Custom Lens do AWS Well-Architected Tool, foram realizadas análises com a ferramenta e Custom Lens de práticas recomendadas. As oportunidades de melhoria geradas pelas análises recebem prioridade nos sprints de software.

Etapas da implementação

1. Realize análises de arquitetura periódicas de sua workload de produção pelo menos anualmente. Use um padrão de arquitetura documentado que inclua práticas recomendadas específicas da AWS.
   a. Recomendamos usar seus próprios padrões definidos internamente para essas análises. Se você não tiver um padrão interno, recomendamos usar a AWS Well-Architected Framework.
   b. Você pode usar o AWS Well-Architected Tool para criar um Custom Lens de suas práticas recomendadas internas e realizar a análise de sua arquitetura.
   c. Os clientes podem entrar em contato com o arquiteto de soluções da AWS para realizar uma Análise da Well-Architected Framework da workload deles.
2. Priorize as oportunidades de melhoria identificadas durante a análise em seu processo de desenvolvimento de software.

Nível de esforço do plano de implementação: baixo. É possível usar a AWS Well-Architected Framework para realizar sua análise de arquitetura anual.

Recursos

Práticas recomendadas relacionadas:

- OPS11-BP02 Executar análise pós-incidente: análise pós-incidente é outro gerador de itens de melhoria. Insira as lições aprendidas em sua lista interna de práticas recomendadas de arquitetura.
- OPS11-BP08 Documentar e compartilhar as lições aprendidas: à medida que você desenvolve suas próprias práticas recomendadas de arquitetura, compartilhe-as em sua organização.

Documentos relacionados:

- AWS Well-Architected Tool: Custom lenses
- AWS Whitepaper sobre Well-Architected: The review process (O processo de revisão)
- Customize Well-Architected Reviews using Custom Lenses and the AWS Well-Architected Tool (Personalizar as Revisões de Well-Architected com o uso de Custom Lenses e o AWS Well-Architected Tool)
- Implementing the AWS Well-Architected Custom Lens lifecycle in your organization (Implementar o ciclo de vida do AWS Well-Architected Custom Lens em sua organização)

Vídeos relacionados:

- Well-Architected Labs - Level 100: Custom Lenses on AWS Well-Architected Tool (Well-Architected Labs: nível 100: Custom Lenses no AWS Well-Architected Tool)

Exemplos relacionados:

- O AWS Well-Architected Tool

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 Realizar o gerenciamento de conhecimento

O gerenciamento de conhecimento ajuda os membros da equipe a encontrar as informações para realizar o trabalho deles. Nas organizações de aprendizagem, as informações são compartilhadas livremente, o que capacita as pessoas. As informações podem ser descobertas ou pesquisadas. As informações são precisas e atualizadas. Os mecanismos existem para criar informações, atualizar informações existentes e arquivar informações desatualizadas. O exemplo mais comum de uma plataforma de gerenciamento de conhecimento é um sistema de gerenciamento de conteúdo como uma wiki.

Resultado desejado:

- Os membros da equipe têm acesso a informações precisas e em tempo hábil.
- As informações podem ser pesquisadas.
- Existem mecanismos para adicionar, atualizar e arquivar informações.

Antipadrões comuns:

- Não há um armazenamento de conhecimento centralizado. Os membros da equipe gerenciam suas próprias notas nas máquinas locais.
- Você tem uma wiki hospedada pela própria empresa, mas nenhum mecanismo para gerenciar informações, resultando em informações desatualizadas.
- Alguém identifica a ausência de informações, mas não há nenhum processo para solicitar a adição delas à wiki da equipe. Essa pessoa adiciona as informações por conta própria, mas deixa de realizar uma etapa, resultando em uma interrupção.

Benefícios do estabelecimento desta prática recomendada:

- Os membros da equipe são capacitados, pois as informações são compartilhadas livremente.
- Os novos membros da equipe passam pelo processo de integração mais rapidamente, pois a documentação está atualizada e pode ser pesquisada.
- As informações são precisas, levam a ações concretas e são enviadas em tempo hábil.

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

Orientações para a implementação

O gerenciamento de conhecimento é uma faceta importante das organizações de aprendizagem. Para começar, é necessário ter um repositório central para armazenar seu conhecimento (como um exemplo comum, uma wiki hospedada pela própria empresa). É necessário desenvolver processos para adicionar, atualizar e arquivar conhecimento. Desenvolva padrões para o que deve ser documentado e permita que todos contribuam.

Exemplo de clientes

A Loja UmaEmpresa hospeda uma wiki interna em que todo o conhecimento é armazenado. Os membros da equipe são incentivados a adicionar informações na base de conhecimento à medida que realizam suas tarefas diárias. Trimestralmente, uma equipe multifuncional avalia quais páginas estão mais desatualizadas e determina se elas devem ser arquivadas ou atualizadas.

Etapas da implementação

1. Comece identificando o sistema de gerenciamento de conteúdo em que o conhecimento será armazenado. Obtenha o consentimento das partes interessadas em sua organização.
    a. Se você não tiver um sistema de gerenciamento de conteúdo, considere desenvolver uma wiki hospedada pela própria empresa ou usar um repositório de controle de versão como ponto de partida.
2. Desenvolva runbooks para adicionar, atualizar e arquivar informações. Instrua a equipe sobre esses processos.
3. Identifique quais conhecimentos devem ser armazenados no sistema de gerenciamento de conteúdo. Comece com atividades diárias (runbooks e manuais) que os membros da equipe realizam. Trabalhe com as partes interessadas para priorizar qual conhecimento deve ser adicionado.
4. Periodicamente, trabalhe com as partes interessadas para identificar informações desatualizadas e arquive-as ou atualize-as.

Nível de esforço do plano de implementação: médio. Se você não tiver um sistema de gerenciamento de conteúdo, defina uma wiki hospedada pela própria empresa ou um repositório de documentos com controle de versão.

Recursos

Práticas recomendadas relacionadas:

- OPS11-BP08 Documentar e compartilhar as lições aprendidas – O gerenciamento de conhecimento facilita o compartilhamento de informações sobre as lições aprendidas.

Documentos relacionados:

- Atlassian: gerenciamento do conhecimento

Exemplos relacionados:

- DokuWiki
- Gollum
- MediaWiki
- Wiki.js

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

O pilar Segurança refere-se à capacidade de proteger dados, sistemas e ativos para utilizar as tecnologias de nuvem para melhorar sua segurança. Você pode encontrar orientações prescritivas sobre implementação no whitepaper Pilar de segurança.

Áreas de práticas recomendadas

- Fundamentos de segurança
- Gerenciamento de identidade e acesso
- Detecção
- Proteção de infraestrutura
- Proteção de dados
- Resposta a incidentes
- Segurança de aplicações

Fundamentos de segurança

Pergunta

- SEGURANÇA 1. Como operar com segurança sua workload?

SEGURANÇA 1. Como operar com segurança sua workload?

Para operar sua workload com segurança, você deve aplicar as práticas recomendadas 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 as propriedades e o usuário raiz das contas
- 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 ameaças e priorizar mitigações com o uso de um modelo de ameaça
- SEC01-BP08 Avaliar e implementar regularmente novos serviços e recursos de segurança

SEC01-BP01 Separar as workloads usando contas

Estabeleça barreiras de proteção e isolamento entre workloads e ambientes (como de produção, desenvolvimento e teste) por meio de uma estratégia de várias contas. A separação em nível de conta é altamente recomendável, pois ela oferece um limite de isolamento robusto para segurança, faturamento e acesso.

Resultado desejado: uma estrutura de conta que isola operações na nuvem, workloads não relacionadas e ambientes em contas separadas, aumentando a segurança em toda a infraestrutura de nuvem.

Antipadrões comuns:

- Colocação de várias workloads não relacionadas com diferentes níveis de confidencialidade na mesma conta.
- Estrutura de unidade organizacional (UO) definida de forma inadequada.

Benefícios do estabelecimento desta prática recomendada:

- Redução do escopo de impacto se uma workload for acessada acidentalmente.
- Governança central de acesso a serviços, recursos e regiões da AWS.
- Manutenção da segurança da infraestrutura de nuvem com políticas e administração centralizada de serviços de segurança.
- Criação de contas automatizada e processo de manutenção.
- Auditoria centralizada da infraestrutura de conformidade e requisitos regulatórios.

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

Orientação de implementação

As Contas da AWS oferecem um limite de isolamento de segurança entre workloads ou recursos que operam em diferentes níveis de confidencialidade. Para utilizar esse limite de isolamento, a AWS oferece ferramentas para gerenciar em grande escala suas workloads de nuvem por meio de uma estratégia de várias contas. Para ter orientações sobre os conceitos, os padrões e a implementação de uma estratégia de várias contas na AWS, consulte Organizar seu ambiente da AWS com o uso de várias contas.

Quando você tem várias Contas da AWS no gerenciamento central, elas devem ser organizadas em uma hierarquia definida por camadas de unidades organizacionais (UOs). Desse modo, os controles de segurança podem ser organizados e aplicados às UOs e às contas membros, estabelecendo controles preventivos consistentes nas contas membros da organização. Os controles de segurança são herdados, permitindo que você filtre as permissões disponíveis para as contas membros localizadas em níveis inferiores de uma hierarquia de UOs. Um bom design aproveita essa herança para reduzir o número e a complexidade das políticas de segurança necessárias para obter os controles de segurança desejados para cada conta membro.

O AWS Organizations e o AWS Control Tower são dois serviços que você pode utilizar para implementar e gerenciar essa estrutura de várias contas em seu ambiente da AWS. O AWS Organizations possibilita organizar as contas em uma hierarquia definida por uma ou mais camadas

de UOs, em que cada UO contém uma série de contas membros. As políticas de controle de serviços (SCPs) permitem que o administrador da organização estabeleça controles preventivos detalhados nas contas membros, e o AWS Config pode ser utilizado para estabelecer controles proativos e de detecção nessas contas. Muitos serviços da AWS integram-se ao AWS Organizations para oferecer controles administrativos delegados e realizar tarefas específicas do serviço em todas as contas membros da organização.

Estruturado sobre o AWS Organizations, o AWS Control Tower oferece práticas recomendadas de um clique para um ambiente da AWS de várias contas com uma zona de pouso. A zona de pouso é o ponto de entrada para o ambiente de várias contas estabelecido pelo Control Tower. O Control Tower oferece vários benefícios em comparação com o AWS Organizations. Três benefícios que oferecem governança aprimorada de contas são:

- Barreiras de proteção de segurança obrigatórias e integradas que são aplicadas automaticamente às contas admitidas na organização.
- Barreiras de proteção opcionais que podem ser ativadas ou desativadas em determinado conjunto de UOs.
- O AWS Control Tower Account Factory oferece implantação automatizada de contas que contêm linhas de base aprovadas e opções de configuração em sua organização.

Etapas da implementação

1. Projetar uma estrutura de unidade organizacional: uma estrutura de unidade organizacional projetada adequadamente reduz o trabalho de gerenciamento necessário para criar e manter políticas de controle de serviços e outros controles de segurança. Sua estrutura de unidade organizacional deve estar alinhada com as necessidades, a confidencialidade dos dados e a estrutura de workload de sua empresa.
2. Criar uma zona de pouso para seu ambiente de várias contas: uma zona de pouso oferece uma base consistente de infraestrutura e segurança na qual sua organização pode desenvolver, executar e implantar workloads com rapidez. É possível usar uma zona de pouso personalizada ou o AWS Control Tower para orquestrar seu ambiente.
3. Estabelecer barreiras de proteção: implemente barreiras de proteção consistentes para seu ambiente por meio da zona de pouso. O AWS Control Tower oferece uma lista de controles obrigatórios e opcionais que podem ser implantados. Os controles obrigatórios são implantados automaticamente na implementação do Control Tower. Leia a lista de controles opcionais e altamente recomendados e implemente controles adequados às suas necessidades.

4. Restringir o acesso a regiões adicionadas recentemente: para novas Regiões da AWS, recursos do IAM, como usuários e perfis, serão propagados somente para as regiões especificadas. Essa ação pode ser realizada por meio do console ao usar o Control Tower ou ajustando as políticas de permissões do IAM no AWS Organizations.
5. Considerar o AWS CloudFormation StackSets: o StackSets ajuda você a implantar recursos, como grupos, políticas e perfis do IAM em diferentes regiões e Contas da AWS por meio de um modelo aprovado.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP04 Contar com um provedor de identidades centralizado

Documentos relacionados:

- AWS Control Tower
- Diretrizes de auditoria de segurança da AWS
- Práticas recomendadas do IAM
- Usar o CloudFormation StackSets para fornecer recursos entre várias regiões e Contas da AWS
- Perguntas frequentes sobre o Organizations
- Terminologia e conceitos do AWS Organizations
- Práticas recomendadas para políticas de controle de serviços do AWS Organizations em um ambiente de várias contas
- Guia de referência de gerenciamento de contas da AWS
- Organização do ambiente usando várias contas da AWS

Vídeos relacionados:

- Habilitar a adoção da AWS em grande escala com automação e governança
- Práticas recomendadas de segurança de acordo com o Well-Architected
- Criar e administrar várias contas com o uso do AWS Control Tower
- Habilitar o Control Tower para organizações existentes

Workshops relacionados:

- Dia de imersão no Control Tower

SEC01-BP02 Proteger as propriedades e o usuário raiz das contas

O usuário raiz é o mais privilegiado de uma Conta da AWS, com acesso administrativo integral a todos os recursos da conta, e em alguns casos não pode ser restringido por políticas de segurança. Desabilitar o acesso programático ao usuário raiz, estabelecer controles apropriados para ele e evitar o uso rotineiro desse usuário ajuda a reduzir o risco de exposição acidental das credenciais raiz e o subsequente comprometimento do ambiente de nuvem.

Resultado desejado: proteger o usuário raiz ajuda a reduzir a chance de danos acidentais ou intencionais decorrentes do mau uso das respectivas credenciais. Estabelecer controles de detecção também pode alertar o pessoal apropriado quando se realizam ações utilizando o usuário raiz.

Antipadrões comuns:

- Utilizar o usuário raiz para outras tarefas que não sejam aquelas que exigem credenciais do usuário raiz.
- Negligenciar os testes dos planos de contingência regularmente a fim de verificar a funcionalidade da infraestrutura, dos processos e dos funcionários essenciais durante uma emergência.
- Considerar apenas o fluxo típico de login de contas e não considerar nem testar métodos de recuperação de contas alternativos.
- Não lidar com DNS, servidores de e-mail e operadoras de telefonia como parte do perímetro de segurança essencial, pois eles são usados no fluxo de recuperação de contas.

Benefícios do estabelecimento desta prática recomendada: proteger o acesso ao usuário raiz cria a confiança de que as ações em sua conta estão controladas e auditadas.

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

Orientação de implementação

A AWS oferece muitas ferramentas para ajudar a proteger sua conta. No entanto, como algumas dessas medidas não estão habilitadas por padrão, é necessário implementá-las diretamente. Leve em consideração essas recomendações como etapas fundamentais para proteger sua Conta da AWS. Ao implementar essas etapas, é importante criar um processo para avaliar e monitorar os controles de segurança de forma contínua.

Ao criar uma Conta da AWS pela primeira vez, você começa com uma identidade que tem acesso completo a todos os recursos e serviços da AWS na conta. Essa identidade é chamada de usuário raiz da Conta da AWS. É possível fazer login como usuário raiz usando o endereço de e-mail e a senha utilizados para criar a conta. Devido ao acesso elevado concedido ao usuário raiz da AWS, é necessário limitar o uso do usuário raiz da AWS à realização de tarefas que o exijam especificamente. As credenciais de login do usuário raiz devem ser bem protegidas, e a autenticação multifator (MFA) sempre deve ser habilitada para o usuário raiz da Conta da AWS.

Além do fluxo de autenticação normal para fazer login com seu usuário raiz usando um nome de usuário, senha e o dispositivo de autenticação multifator (MFA), há fluxos de recuperação de contas para fazer login com seu usuário raiz da Conta da AWS com o endereço de e-mail e o número de telefone associados à sua conta. Dessa forma, é igualmente importante proteger a conta de e-mail do usuário raiz para a qual o e-mail de recuperação é enviado e o número de telefone associado à conta. Além disso, considere possíveis dependências circulares em que o endereço de e-mail associado ao usuário raiz é hospedado em servidores de e-mail ou recursos de serviço de nome de domínio (DNS) da mesma Conta da AWS.

Ao usar o AWS Organizations, há várias Contas da AWS, e cada uma tem um usuário raiz. Uma conta é designada como a conta de gerenciamento e várias camadas de contas membros podem ser adicionadas à conta de gerenciamento. Priorize a proteção do usuário raiz de sua conta de gerenciamento e, depois, os usuários raiz das contas membros. A estratégia para proteger o usuário raiz de sua conta de gerenciamento pode diferir da utilizada nos usuários raiz de suas contas membros, e é possível implementar controles de segurança preventivos nos usuários raiz dessas contas.

Etapas da implementação

As etapas de implementação a seguir são recomendadas para estabelecer controles para o usuário raiz. Quando aplicável, as recomendações têm referências cruzadas com o Benchmark do CIS AWS Foundations versão 1.4.0. Além dessas etapas, consulte as Diretrizes de práticas recomendadas da AWS para proteger os recursos e a Conta da AWS.

Controles preventivos

1. Configure informações de contato precisos para a conta.
    a. Essas informações são usadas para o fluxo de recuperação de senha perdida, o fluxo de recuperação de conta de dispositivo MFA perdida e para comunicações com sua equipe sobre segurança crítica.

   b. Utilize um endereço de e-mail hospedado por seu domínio corporativo, preferencialmente uma lista de distribuição, como o endereço de e-mail do usuário raiz. O uso de uma lista de distribuição em vez da conta de e-mail de um indivíduo oferece redundância e continuidade adicionais para o acesso à conta raiz por longos períodos.

   c. O número de telefone listado nas informações de contato deve ser um telefone dedicado e seguro para esse fim. O número de telefone não deve ser listado nem compartilhado com ninguém.

2. Não crie chaves de acesso para o usuário raiz. Se houver chaves de acesso, remova-as (CIS 1.4).

   a. Elimine todas as credenciais programáticas de longa duração (chaves de acesso e secretas) para o usuário raiz.

   b. Se já houver chaves de acesso do usuário raiz, será necessário fazer a transição dos processos que utilizam essas chaves para utilizar chaves de acesso temporárias de um perfil do AWS Identity and Access Management (IAM), depois excluir as chaves de acesso do usuário raiz.

3. Determine se você precisa armazenar credenciais para o usuário raiz.

   a. Ao usar o AWS Organizations para criar contas membros, a senha inicial do usuário raiz em novas contas membros é definida como um valor aleatório que não é exposto a você. Se necessário, considere utilizar o fluxo de redefinição de senha de sua conta de gerenciamento do AWS Organizations para obter acesso à conta membro.

   b. Para Contas da AWS autônomas ou a conta de gerenciamento do AWS Organizations, considere criar e armazenar de forma segura as credenciais do usuário raiz. Ativar a MFA para o usuário raiz

4. Ative os controles preventivos para os usuários raiz das contas membros em ambientes de várias contas da AWS.

   a. Considere habilitar a barreira de proteção preventiva Desautorizar criação de chaves de acesso raiz para o usuário raiz para contas membros.

   b. Considere habilitar a barreira de proteção preventiva Desautorizar criação como um usuário raiz para contas membros.

5. Se você precisar de credenciais para o usuário raiz:

   a. Use uma senha complexa.

   b. Ative a autenticação multifator (MFA) para o usuário raiz, especialmente para contas (pagantes) de gerenciamento do AWS Organizations (CIS 1.5).

c. Considere o uso de dispositivos de MFA de hardware para ter resiliência e segurança, pois os dispositivos de uso único reduzem as chances de os dispositivos que contêm seus códigos de MFA serem reutilizados para outros fins. Garanta que os dispositivos de MFA de hardware alimentados por bateria sejam substituídos regularmente. (CIS 1.6)

   - Para configurar a MFA para o usuário raiz, siga as instruções para habilitar uma MFA virtual ou um dispositivo de MFA de hardware.

d. Considere registrar vários dispositivos de MFA para backup. Até oito dispositivos de MFA são permitidos por conta.

   - Registrar mais de um dispositivo de MFA para o usuário raiz desabilita automaticamente o fluxo para recuperar sua conta se o dispositivo de MFA for perdido.

e. Armazene a senha com segurança e considere as dependências circulares se for armazenar a senha eletronicamente. Não armazene a senha de uma forma que exija o acesso à mesma Conta da AWS para obtê-la.

6. Opcional: considere estabelecer um cronograma de alternância de senha periódica para o usuário raiz.

   - As práticas recomendadas de gerenciamento de credenciais dependem de seus requisitos regulatórios e de política. Os usuários raiz protegidos por MFA não dependem da senha como um único fator de autenticação.
   - A alteração periódica da senha de usuário raiz reduz o risco de mau uso de uma senha exposta acidentalmente.

Controles de detecção

- Crie alarmes para detectar o uso das credenciais raiz (CIS 1.7). Quando habilitado, o Amazon GuardDuty monitora e alerta o uso de credenciais da API do usuário raiz por meio da descoberta RootCredentialUsage.
- Avalie e implemente os controles de detecção incluídos no pacote de conformidade do Pilar de segurança do AWS Well-Architected para AWS Config ou, se usar o AWS Control Tower, os controles altamente recomendados disponíveis no Control Tower.

Orientação operacional

- Determine quem na organização deve ter acesso às credenciais do usuário raiz.

    - Use uma regra de duas pessoas de forma que um indivíduo tenha acesso a todas as credenciais necessárias e MFA para obter acesso de usuário raiz.
    - Verifique se é a organização, e não um único indivíduo, que mantém controle sobre o número de telefone e alias de e-mail associados à conta (que são utilizados para redefinição de senha e fluxo de redefinição de MFA).
- Utilize o usuário raiz apenas como uma exceção (CIS 1.7).
    - O usuário raiz da AWS não deve ser usado para tarefas diárias, mesmo que sejam tarefas administrativas. Somente faça login como usuário raiz para realizar tarefas da AWS que o exijam especificamente. Todas as outras ações devem ser realizadas por outros usuários que assumem perfis apropriados.
- Confira periodicamente se o acesso ao usuário raiz está funcionando de forma que os procedimentos sejam testados antes de uma situação de emergência que exija o uso das credenciais do usuário raiz.
- Confira periodicamente se o endereço de e-mail associado à conta e os listados em Contatos alternativos funcionam. Monitore as caixas de entrada de e-mail das quais você recebe notificações de segurança <abuse@amazon.com>. Além disso, garanta que todos os números de telefone associados à conta estejam funcionando.
- Prepare um procedimento de resposta a incidentes para responder ao mau uso da conta raiz. Consulte o Guia de resposta a incidentes de segurança da AWS e as práticas recomendadas na seção “Resposta a incidentes” do whitepaper Pilar Segurança para ter mais informações sobre como criar uma estratégia de resposta a incidentes para sua Conta da AWS.

Recursos

Práticas recomendadas relacionadas:

- SEC01-BP01 Separar as workloads usando contas
- SEC02-BP01 Usar mecanismos de login fortes
- SEC03-BP02 Conceder acesso com privilégio mínimo
- SEC03-BP03 Estabelecer processo de acesso de emergência
- SEC10-BP05 Acesso pré-provisionado

Documentos relacionados:

- AWS Control Tower

- Diretrizes de auditoria de segurança da AWS
- Práticas recomendadas do IAM
- Amazon GuardDuty: alerta de uso de credenciais raiz
- Orientações passo a passo sobre como monitorar o uso de credenciais raiz por meio do CloudTrail
- Tokens de MFA aprovados para uso com a AWS
- Implementar o acesso de quebra de vidro na AWS
- Os dez principais itens de segurança para aprimorar sua Conta da AWS
- O que fazer se eu notar atividade não autorizada em minha Conta da AWS?

Vídeos relacionados:

- Habilitar a adoção da AWS em grande escala com automação e governança
- Práticas recomendadas de segurança de acordo com o Well-Architected
- Limitar o uso de credenciais raiz da AWS do AWS re:inforce 2022: Práticas recomendadas de segurança com o AWS IAM

Exemplos e laboratórios relacionados:

- Laboratório: Conta da AWS e usuário raiz

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 ameaças e priorizar mitigações com o uso de um modelo de ameaça

Realize a modelagem de ameaças para identificar e manter um registro atualizado de possíveis ameaças e mitigações associadas para sua workload. Priorize suas ameaças e adapte as mitigações de controles de segurança para prevenir, detectar e responder. Revise e mantenha isso no contexto de sua workload e no cenário de segurança em evolução.

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

Orientações para a implementação

O que é modelagem de ameaças?

“A modelagem de ameaças serve para identificar, comunicar e compreender as ameaças e as mitigações no contexto de proteção de algo de valor.” The Open Web Application Security Project (OWASP) Application Threat Modeling

Por que você deve ter um modelo de ameaças?

Os sistemas são complexos e se tornam cada vez mais intrincados e qualificados com o passar do tempo, oferecendo maior valor empresarial e maior satisfação e engajamento do cliente. Isso significa que as decisões de design de TI precisam considerar um número cada vez maior de casos de uso. Essa complexidade e o número de permutações de caso de uso geralmente tornam as abordagens não estruturadas ineficazes para encontrar e mitigar ameaças. Em vez disso, você precisa de uma abordagem sistemática para enumerar as possíveis ameaças ao sistema, elaborar mitigações e priorizá-las a fim de garantir que os recursos limitados de sua organização tenham impacto máximo na melhoria do procedimento geral de segurança do sistema.

A modelagem de ameaças foi projetada para oferecer essa abordagem sistemática, com o objetivo de encontrar e resolver problemas na fase inicial do processo de design, quando as mitigações têm custo e esforço relativamente baixos em comparação com a fase posterior do ciclo de vida. Essa abordagem está alinhada ao princípio de segurança shift left (mover para a esquerda) do setor. Por fim, a modelagem de ameaças é integrada ao processo de gerenciamento de riscos de uma organização e ajuda a impulsionar as decisões sobre quais controles implementar usando uma abordagem orientada a ameaças.

Quando a modelagem de ameaças deve ser realizada?

Inicie a modelagem de ameaças o quanto antes no ciclo de vida de sua workload. Isso oferece a você maior flexibilidade sobre o que fazer com as ameaças identificadas. Muito semelhante aos bugs de software, quanto mais cedo você identificar ameaças, mais econômico será resolvê-las. Um modelo de ameaças é um documento ativo e deve continuar a evoluir à medida que suas workloads mudam. Revise seus modelos de ameaça no decorrer do tempo, inclusive quando há uma alteração importante ou uma alteração no cenário de ameaças ou ao adotar um novo recurso ou serviço.

Etapas da implementação

Como podemos realizar a modelagem de ameaças?

Há muitas formas diferentes de realizar a modelagem de ameaças. Muito semelhante às linguagens de programação, há vantagens e desvantagens em cada uma, e é necessário escolher a forma mais adequada para você. Uma abordagem é começar com o Shostack's 4 Question Frame for Threat Modeling (Estrutura de quatro perguntas do Shostack para modelagem de ameaças), que apresenta perguntas abertas a fim de oferecer estrutura ao seu exercício de modelagem de ameaças:

1. Em que você está trabalhando?

   A finalidade dessa pergunta é ajudar você a entender e chegar a um acordo sobre o sistema que você está construindo e os detalhes sobre ele que são relevantes para a segurança. A criação de

um modelo ou um diagrama é a forma mais comum de responder a essa pergunta, pois ele ajuda você a visualizar o que você está construindo; por exemplo, usando um fluxograma de dados. Escrever as suposições e os detalhes importantes sobre seu sistema também ajuda a definir o que está no escopo. Isso permite que todos que estão contribuindo para o modelo de ameaças se concentrem na mesma coisa e evitem desvios demorados para tópicos fora do escopo (inclusive versões desatualizadas do sistema). Por exemplo, se você estiver criando uma aplicação web, provavelmente não vale a pena criar uma modelagem de ameaças da sequência de inicialização confiável do sistema operacional para clientes de navegador, pois não há nenhuma possibilidade de seu design ter influência nisso.

2. O que pode dar errado?

   É nessa fase que você identifica ameaças ao seu sistema. Ameaças são ações ou eventos acidentais ou intencionais que têm impactos indesejados que podem afetar a segurança de seu sistema. Sem um claro entendimento do que pode dar errado, não há o que fazer sobre isso.

   Não há uma lista canônica do que pode dar errado. A criação dessa lista exige etapas de brainstorming e colaboração entre todas as pessoas de sua equipe e as pessoas relevantes envolvidas no exercício de modelagem de ameaças. Você pode auxiliar suas etapas de brainstorming utilizando um modelo para identificar ameaças, como o STRIDE, que sugere categorias diferentes para avaliar: Spoofing (Falsificação), Tampering (Violação), Repudiation (Repúdio), Information disclosure (Divulgação de informações), Denial of service (Negação de serviço) e Elevation of privilege (Elevação de privilégio). Além disso, talvez você queira auxiliar as etapas de brainstorming revisando as listas existentes e a pesquisa para inspiração, como o OWASP Top 10, o HiTrust Threat Catalog e o catálogo de ameaças de sua própria organização.

3. O que estamos fazendo a respeito?

   Como no caso da primeira pergunta, não há uma lista canônica de todas as mitigações possíveis. A entradas nessa etapa são as ameaças identificadas, as pessoas e as áreas de melhoria da etapa anterior.

   Segurança e conformidade são responsabilidades compartilhadas entre você e a AWS. É importante entender que ao perguntar “O que vamos fazer a respeito?” você também pergunte “Quem é responsável por fazer algo a respeito?”. Entender o equilíbrio entre suas responsabilidades e as da AWS ajuda a definir o escopo de seu exercício de modelagem de ameaças para as mitigações que estão sob seu controle, que, geralmente, são uma combinação de opções de configuração de serviços da AWS e suas mitigações específicas ao sistema.

No que se refere à responsabilidade compartilhada da AWS, você descobrirá que os serviços da AWS estão no escopo de muitos programas de conformidade. Esses programas ajudam você a entender os controles sólidos implementados na AWS para manter a segurança e a conformidade da nuvem. Os relatórios de auditoria desses programas estão disponíveis para download para clientes da AWS do AWS Artifact.

Seja quais forem os serviços da AWS que você esteja utilizando, sempre há um elemento de responsabilidade do cliente, e as mitigações alinhadas a essas responsabilidades devem ser incluídas em seu modelo de ameaças. Para mitigações de controle de segurança dos próprios serviços da AWS, convém considerar a implementação de controles de segurança em todos os domínios; por exemplo, domínios como gerenciamento de identidade e acesso (autenticação e autorização), proteção de dados (em repouso e em trânsito), segurança de infraestrutura, registro em log e monitoramento. A documentação de cada serviço da AWS tem um capítulo dedicado à segurança que oferece orientações sobre os controles de segurança a serem considerados como mitigações. É importante considerar o código que você está escrevendo e suas dependências e pensar nos controles que você poderia implementar para resolver essas ameaças. Esses controles podem ser fatores como validação de entrada, processamento de sessões e processamento de limites. Com frequência, a maioria das vulnerabilidades é introduzida em código personalizado, então concentre-se nessa área.

4. Fizemos um bom trabalho?

   O objetivo é a sua equipe e a organização aprimorarem a qualidade dos modelos de ameaças e a velocidade na qual você está realizando a modelagem de ameaças no decorrer do tempo. Essas melhorias vêm de uma combinação de prática, aprendizado, instrução e revisão. Para se aprofundar e trabalhar, é recomendável que você e sua equipe concluam o curso de treinamento Threat modeling the right way for builders (Modelagem de ameças da maneira certa para desenvolvedores) ou o respectivo workshop. Além disso, se você estiver procurando orientações sobre como integrar a modelagem de ameaças ao ciclo de vida do desenvolvimento de aplicações da organização, consulte a publicação Como abordar a modelagem de ameaças no Blog de segurança da AWS.

Threat Composer

Para ajudar e fornecer orientações ao criar a modelagem de ameaças, considere usar a ferramenta Threat Composer, que visa reduzir o tempo de maturação na modelagem de ameaças. Essa ferramenta ajuda a fazer o seguinte:

- Escrever declarações úteis sobre ameaças alinhadas à gramática de ameaças que funcionem em um fluxo de trabalho natural e não linear.
- Gerar um modelo de ameaça legível por humanos.
- Gerar um modelo de ameaça legível por máquina para permitir tratar os modelos de ameaças como código.
- Ajudar a identificar rapidamente as áreas de melhoria da qualidade e de cobertura usando o painel do Insights.

Para obter mais referências, acesse o Threat Composer e alterne para o Example Workspace definido pelo sistema.

Recursos

Práticas recomendadas relacionadas:

- 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-BP08 Avaliar e implementar regularmente novos serviços e recursos de segurança

Documentos relacionados:

- Como abordar a modelagem de ameaças (Blog de segurança da AWS)
- NIST: Guia para modelagem de ameaças de sistemas centrados em dados

Vídeos relacionados:

- AWS Summit ANZ 2021: Como abordar a modelagem de ameaças
- AWS Summit ANZ 2022: Escalar a segurança: otimizar para ter uma entrega rápida e segura

Treinamento relacionado:

- Threat modeling the right way for builders (Modelagem de ameaças da maneira certa para desenvolvedores): treinamento autoguiado virtual do AWS Skill Builder
- Threat modeling the right way for builders – AWS Workshop (Modelagem de ameaças da maneira certa para desenvolvedores)

Ferramentas relacionadas:

- Threat Composer

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

- SEGURANÇA 2. Como gerenciar a autenticação de pessoas e máquinas?
- SEGURANÇA 3. Como gerenciar permissões para pessoas e máquinas?

SEGURANÇA 2. Como gerenciar 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 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: aplicações de serviço, ferramentas operacionais e workloads precisam de uma identidade para fazer solicitações a serviços da AWS, como para ler dados. 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 alternância periódica das credenciais
- SEC02-BP06 Utilizar grupos e atributos de usuários

SEC02-BP01 Usar mecanismos de login fortes

Os logins (autenticação com credenciais de login) podem apresentar riscos quando não são usados mecanismos, como autenticação multifator (MFA), especialmente em situações em que as credenciais de login foram divulgadas acidentalmente ou podem ser deduzidas com facilidade. Utilize mecanismos de login fortes para reduzir essas riscos exigindo MFA e políticas de senhas fortes.

Resultado desejado: reduzir os riscos de acesso acidental a credenciais na AWS usando mecanismos de login fortes para usuários do AWS Identity and Access Management (IAM), o usuário raiz da Conta da AWS, o AWS IAM Identity Center (sucessor do AWS Single Sign-On), e provedores de identidades de terceiros. Isso significa exigir MFA, impor políticas de senhas fortes e detectar comportamento de login anômalo.

Antipadrões comuns:

- Não impor uma política de senhas fortes para suas identidades incluindo senhas complexas e MFA.
- Compartilhar as mesmas credenciais entre usuários diferentes.
- Não utilizar controles de detecção para logins suspeitos.

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

Orientação de implementação

Há muitas formas de identidades humanas fazerem login na AWS. É prática recomendada da AWS depender de um provedor de identidades centralizado utilizando federação (federação direta ou usando AWS IAM Identity Center) ao realizar a autenticação na AWS. Nesse caso, você deve estabelecer um processo de login seguro com seu provedor de identidades ou o Microsoft Active Directory.

Ao abrir pela primeira vez uma Conta da AWS, você começa com um usuário raiz da Conta da AWS. Você só deve usar o usuário raiz da conta para configurar o acesso para seus usuários (e para tarefas que exijam o usuário raiz. É importante ativar a MFA para o usuário raiz da conta logo após abrir sua Conta da AWS e para proteger o usuário raiz usando o Guia de práticas recomendadas da AWS.

Se você criar usuários no AWS IAM Identity Center, proteja o processo de login nesse serviço. Para identidades dos consumidores, é possível usar o Amazon Cognito user pools e proteger o processo

de login nesse serviço ou usar os provedores de identidades compatíveis com o Amazon Cognito user pools.

Se estiver usando usuários do AWS Identity and Access Management (IAM), você protegerá o processo de login com o IAM.

Seja qual for o método de login, é essencial impor uma política de login forte.

Etapas da implementação

Veja a seguir as recomendações gerais de login forte. As configurações reais devem ser definidas pela política de sua empresa ou usando um padrão como NIST 800-63.

- Exija MFA. É prática recomendada IAM exigir MFA para identidades humanas e workloads. A ativação da MFA oferece uma camada adicional de segurança que exige que os usuários forneçam credenciais de login e uma senha de uso único (OTP) ou uma string gerada e verificada criptograficamente por um dispositivo de hardware.
- Imponha um comprimento mínimo de senha, que é um fator essencial da força da senha.
- Imponha complexidade para tornar as senhas mais difíceis de deduzir.
- Permita que os usuários alterem suas próprias senhas.
- Crie identidades individuais em vez de credenciais compartilhadas. Com a criação de identidades individuais, é possível fornecer a cada usuário um conjunto exclusivo de credenciais de segurança. Os usuários individuais oferecem a capacidade de auditar a atividade de cada usuário.

Recomendações do IAM Identity Center:

- O IAM Identity Center oferece uma política de senha predefinida ao usar o diretório padrão que estabelece o comprimento da senha, a complexidade e requisitos de reutilização.
- Ative a MFA e defina a configuração de reconhecimento de contexto ou sempre ativo da MFA quando a origem da identidade for o diretório padrão, o AWS Managed Microsoft AD ou o AD Connector.
- Permita que os usuários registrem seus próprios dispositivos de MFA.

Recomendações de diretório do Amazon Cognito user pools:

- Defina as configurações de força da senha.
- Exija MFA dos usuários.

- Use as configurações de segurança avançadas do Amazon Cognito user pools para recursos como autenticação adaptável que podem bloquear logins suspeitos.

Recomendações para usuários do IAM:

- Em teoria, você está utilizando IAM Identity Center ou federação direta. No entanto, talvez você precise de usuários do IAM. Nesse caso, defina uma política de senha para usuários do IAM. A política de senhas pode ser usada para definir requisitos como extensão mínima ou a obrigatoriedade de uso de caracteres não alfabéticos.
- Crie uma política do IAM com o objetivo de impor login de MFA para que os usuários possam gerenciar suas próprias senhas e dispositivos de MFA.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP03 Armazenar e usar segredos com segurança
- SEC02-BP04 Contar com um provedor de identidades centralizado
- SEC03-BP08 Compartilhar recursos com segurança em sua organização

Documentos relacionados:

- Política de senha do AWS IAM Identity Center (sucessor do AWS Single Sign-On)
- Política de senha do usuário do IAM
- Definir a senha do usuário raiz da Conta da AWS
- Política de senha do Amazon Cognito
- Credenciais da AWS
- Práticas recomendadas de segurança no IAM

Vídeos relacionados:

- Gerenciar permissões de usuário em grande escala com o AWS IAM Identity Center
- Dominar a identidade em todos os aspectos

SEC02-BP02 Usar credenciais temporárias

Ao realizar qualquer tipo de autenticação, é melhor utilizar credenciais temporárias em vez de credenciais de longo prazo a fim de reduzir ou eliminar riscos, como credenciais que são divulgadas acidentalmente, compartilhadas ou roubadas.

Resultado desejado: para reduzir o risco de credenciais de longo prazo, use credenciais temporárias sempre que possível para identidades humanas e de máquina. Credenciais de longo prazo criam muitos riscos, por exemplo, é possível fazer upload delas em código para repositórios públicos do GitHub. Ao utilizar credenciais temporárias, você reduz significativamente as chances de comprometimento das credenciais.

Antipadrões comuns:

- Desenvolvedores que usam chaves de acesso de longo prazo de IAM users em vez de obter credenciais temporárias da CLI usando federação.
- Desenvolvedores que incorporam chaves de acesso de longo prazo no código e fazem upload desse código para repositórios públicos do Git.
- Desenvolvedores que incorporam chaves de acesso de longo prazo em aplicações móveis que, depois, são disponibilizadas em lojas de aplicações.
- Usuários que compartilham chaves de acesso de longo prazo com outros usuários ou funcionários que deixam a empresa e não devolvem as chaves de acesso de longo prazo.
- Utilizar chaves de acesso de longo prazo para identidades de máquina quando é possível usar credenciais temporárias.

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

Orientação de implementação

Utilize credenciais de segurança temporárias em vez de credenciais de longo prazo para todas as solicitações da AWS CLI e API. As solicitações de API e CLI para serviços da AWS devem, em quase todos os casos, ser assinadas com chaves de acesso da AWS. Essas solicitações podem ser assinadas com credenciais temporárias ou de longo prazo. A única vez que você deve usar credenciais de longo prazo, também conhecidas como chaves de acesso de longo prazo, é se você estiver utilizando um usuário do IAM ou o usuário raiz da Conta da AWS. Quando você usa federação na AWS ou assume um perfil do IAM por outros métodos, são geradas credenciais temporárias. Mesmo quando você acessa o AWS Management Console utilizando credenciais de

login, credenciais temporárias são geradas para você fazer chamadas para serviços da AWS. Há poucas situações nas quais você precisa de credenciais de longo prazo, e é possível realizar quase todas as tarefas usando credenciais temporárias.

Evitar o uso de credenciais de longo prazo em favor de credenciais temporárias deve andar lado a lado com uma estratégia de reduzir o uso de usuários do IAM em favor da federação e de perfis do IAM. Embora usuários do IAM tenham sido usados para identidades humanas e de máquina no passado, agora recomendamos não utilizá-los para evitar os riscos de utilizar chaves de acesso de longo prazo.

Etapas da implementação

Para identidades humanas, como funcionários, administradores, desenvolvedores, operadores e clientes:

- Você deve contar com um provedor de identidades centralizado e exigir que usuários humanos usem federação com um provedor de identidades para acessar a AWS utilizando credenciais temporárias. A federação para seus usuários pode ser realizada com federação direta para cada Conta da AWS ou usando o AWS IAM Identity Center (sucessor do AWS IAM Identity Center) e o provedor de identidades de sua escolha. A federação oferece uma série de vantagens em comparação com a utilização de usuários do IAM além de eliminar credenciais de longo prazo. Seus usuários também podem solicitar credenciais temporárias da linha de comando para federação direta ou utilizando o IAM Identity Center. Isso significa que há poucos casos de uso que exigem usuários do IAM ou credenciais de longo prazo para seus usuários.
- Ao conceder acesso a recursos em sua Conta da AWS a terceiros, como provedores de software como serviço (SaaS), você pode utilizar perfis entre contas e políticas baseadas em recursos.
- Se você precisar conceder a aplicações de consumidores ou clientes acesso aos seus recursos da AWS, você pode utilizar grupos de identidade do Amazon Cognito ou Amazon Cognito user pools para fornecer credenciais temporárias. As permissões para as credenciais são configuradas por meio de perfis do IAM. Você também pode definir um perfil do IAM separado com permissões limitadas para usuários convidados que não são autenticados.

Para identidades de máquina, talvez seja necessário utilizar credenciais de longo prazo. Nesses casos, você deve exigir que as workloads utilizem credenciais temporárias com perfis da IAM para acessar a AWS.

- Para Amazon Elastic Compute Cloud (Amazon EC2), é possível usar perfis do Amazon EC2.

- O AWS Lambda permite configurar um perfil de execução do Lambda para conceder permissões de serviço a fim de executar ações da AWS usando credenciais temporárias. Há muitos outros modelos semelhantes para os serviços da AWS concederem credenciais temporárias utilizando perfis do IAM.
- Para serviços de IoT, é possível usar o provedor de credenciais de AWS IoT Core para solicitar credenciais temporárias.
- Para sistemas on-premises ou sistemas executados fora da AWS que precisem acessar os recursos da AWS, é possível utilizar o IAM Roles Anywhere.

Há cenários em que credenciais temporárias não são uma opção e talvez seja necessário usar credenciais de longo prazo. Nessas situações, faça auditoria e alterne as credenciais periodicamente e alterne as chaves de acesso regularmente para casos de uso que exijam credenciais de longo prazo. Alguns exemplos que podem exigir credenciais de longo prazo incluem plug-ins do WordPress e clientes da AWS de terceiros. Em situações em que você precisa utilizar credenciais de longo prazo ou para credenciais que não sejam chaves de acesso da AWS, como logins de banco de dados, é possível usar um serviço projetado para lidar com o gerenciamento de segredos, como AWS Secrets Manager. O Secrets Manager torna simples gerenciar, alternar e armazenar com segurança segredos criptografados utilizando serviços compatíveis. Para ter mais informações sobre a alternância de credenciais de longo prazo, consulte Alternar chave de acesso.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP03 Armazenar e usar segredos com segurança
- SEC02-BP04 Contar com um provedor de identidades centralizado
- SEC03-BP08 Compartilhar recursos com segurança em sua organização

Documentos relacionados:

- Credenciais de segurança temporárias
- Credenciais da AWS
- Práticas recomendadas de segurança no IAM
- Perfis do IAM
- IAM Identity Center

- Provedores de identidades e federação
- Alternar chaves de acesso
- Soluções de parceiros de segurança: acesso e controle de acesso
- Usuário raiz da Conta da AWS

Vídeos relacionados:

- Gerenciar permissões de acesso em grande escala com o AWS IAM Identity Center (sucessor do AWS IAM Identity Center
- Dominar a identidade em todos os aspectos

SEC02-BP03 Armazenar e usar segredos com segurança

Uma workload exige um recurso automatizado para comprovar a identidade dela em bancos de dados, recursos e serviços de terceiros. Isso é realizado com o uso de credenciais de acesso secretas, como chaves de acesso de API, senhas e tokens do OAuth. Utilizar um serviço com propósito específico para armazenar, gerenciar e alternar essas credenciais ajuda a reduzir a probabilidade de comprometimento dessas credenciais.

Resultado desejado: implementar um mecanismo para gerenciar com segurança credenciais de aplicações que concretize os seguintes objetivos:

- Identificar quais segredos são necessários para a workload.
- Reduzir o número de credenciais de longo prazo necessárias substituindo-as por credenciais de curto prazo quando possível.
- Estabelecer um armazenamento seguro e uma alternância automatizada das credenciais de longo prazo restantes.
- Auditar o acesso aos segredos existentes na workload.
- Monitorar continuamente para confirmar que nenhum segredo seja incorporado a código-fonte durante o processo de desenvolvimento.
- Reduzir a probabilidade de divulgação acidental de credenciais.

Antipadrões comuns:

- Ausência de alternância de credenciais.

- Armazenar credenciais de longo prazo em código-fonte ou arquivos de configuração.
- Armazenar credenciais em repouso não criptografadas.

Benefícios do estabelecimento desta prática recomendada:

- Os segredos são armazenados com criptografia em repouso e em trânsito.
- O acesso às credenciais é fechado por meio de uma API (pense nisso como uma máquina automática de venda de credenciais).
- O acesso a uma credencial (de leitura e gravação) é auditado e registrado.
- Separação de preocupações: a alternância de credenciais é realizada por um componente separado, que pode ser segregado do restante da arquitetura.
- Os segredos são automaticamente distribuídos sob demanda em componentes de software e a alternância ocorre em um local central.
- O acesso às credenciais pode ser controlado de forma detalhada.

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

Orientação de implementação

Antes, as credenciais usadas para realizar a autenticação em bancos de dados, APIs de terceiros, tokens e outros segredos podiam ser incorporadas em código-fonte ou em arquivos do ambiente. A AWS oferece vários mecanismos para armazenar essas credenciais com segurança, alterná-las automaticamente e auditar o uso delas.

A melhor forma de abordar o gerenciamento de segredos é seguir as orientações de remover, substituir e alternar. A credencial mais segura é a que você não precisa armazenar, gerenciar nem processar. Pode haver credenciais que não sejam mais necessárias ao funcionamento da workload que podem ser removidas com segurança.

Para credenciais que ainda são necessárias ao funcionamento adequado da workload, pode haver uma oportunidade de substituir uma credencial de longo prazo por uma credencial temporária ou de curto prazo. Por exemplo, em vez de codificar uma chave de acesso secreta da AWS, pense em substituir essa credencial de longo prazo por uma temporária utilizando perfis do IAM.

Alguns segredos duradouros podem não ser removidos ou substituídos. Esses segredos podem ser armazenados em um serviço, como o AWS Secrets Manager, no qual eles podem ser armazenados centralmente, gerenciados e alternados regularmente.

Uma auditoria do código-fonte da workload e os arquivos de configuração podem revelar muitos tipos de credencial. A seguinte tabela resume as estratégias para lidar com tipos comuns de credenciais:

Credential type	Description	Suggested strategy
IAM access keys	AWS IAM access and secret keys used to assume IAM roles inside of a workload	Replace: Use Perfis do IAM assigned to the compute instances (such as Amazon EC2 or AWS Lambda) instead. For interoperability with third parties that require access to resources in your Conta da AWS, ask if they support Acesso entre contas da AWS. For mobile apps, consider using temporary credentials through Grupos de identidades (identidades federadas) do Amazon Cognito. For workloads running outside of AWS, consider IAM Roles Anywhere or AWS Systems Manager Hybrid Activations.
SSH keys	Secure Shell private keys used to log into Linux EC2 instances, manually or as part of an automated process	Replace: Use AWS Systems Manager or EC2 Instance Connect to provide programmatic and human access to EC2 instances using IAM roles.
Application and database credentials	Passwords – plain text string	Rotate: Store credentials in AWS Secrets Manager and establish automated rotation if possible.
Amazon RDS and Aurora Admin Database credentials	Passwords – plain text string	Replace: Use the Integração do Secrets Manager ao

Credential type	Description	Suggested strategy
		Amazon RDS or Amazon Aurora. In addition, some RDS database types can use IAM roles instead of passwords for some use cases (for more detail, see Autenticação de banco de dados do IAM).
OAuth tokens	Secret tokens – plain text string	Rotate: Store tokens in AWS Secrets Manager and configure automated rotation.
API tokens and keys	Secret tokens – plain text string	Rotate: Store in AWS Secrets Manager and establish automated rotation if possible.

Um antipadrão comum é incorporar chaves de acesso do IAM ao código-fonte, a arquivos de configuração ou aplicativos móveis. Quando uma chave de acesso do IAM é necessária para comunicação com um serviço da AWS, utilize credenciais de segurança temporárias (de curto prazo). Essas credenciais de curto prazo podem ser fornecidas por meio de perfis do IAM para instâncias do EC2, perfis de execução para funções do Lambda, perfis do Cognito IAM para acesso de usuários móveis e políticas do IoT Core para dispositivos IoT. Ao fazer interface com terceiros, prefira delegar o acesso a um perfil do IAM com o acesso necessário aos recursos de sua conta em vez de configurar um usuário do IAM e enviar a terceiros a chave de acesso secreta desse usuário.

Há muitos casos em que a workload exige o armazenamento de segredos necessários para interoperar com outros serviços e recursos. O AWS Secrets Manager foi concebido especificamente para gerenciar com segurança essas credenciais, bem como o armazenamento, o uso e a alternância de tokens de API, senhas e outras credenciais.

O AWS Secrets Manager oferece cinco recursos principais para proteger o armazenamento e o processamento de credenciais sigilosas: criptografia em repouso, criptografia em trânsito, auditoria abrangente, controle de acesso detalhado e alternância de credenciais extensíveis. Outros serviços de gerenciamento de segredos de parceiros da AWS ou soluções desenvolvidas localmente que oferecem recursos e garantias semelhantes também são aceitáveis.

Etapas da implementação

1. Identifique caminhos de código que contenham credenciais codificadas usando ferramentas automatizadas, como o Amazon CodeGuru.
   - Utilize o Amazon CodeGuru para verificar seus repositórios de código. Depois de concluir a revisão, filtre `Type=Secrets` no CodeGuru para encontrar linhas de código problemáticas.
2. Identifique credenciais que possam ser removidas ou substituídas.
   a. Identifique credenciais não mais necessárias e marque-as para remoção.
   b. Para chaves secretas da AWS incorporadas ao código-fonte, substitua-as por perfis do IAM associados aos recursos necessários. Se parte de sua workload estiver fora do AWS, mas exigir credenciais do IAM para acessar recursos da AWS, considere o IAM Roles Anywhere ou o AWS Systems Manager Hybrid Activations.
3. Para outros terceiros, segredos duradouros que exijam o uso da estratégia de alternância, integre o Secrets Manager ao seu código para recuperar segredos de terceiros no tempo de execução.
   a. O console do CodeGuru pode criar um segredo automaticamente no Secrets Manager utilizando as credenciais descobertas.
   b. Integre a recuperação de segredos do Secrets Manager ao código de sua aplicação.
      - Funções do Lambda Sem Servidor podem usar uma extensão do Lambda independente de linguagem.
      - Para instâncias ou contêineres do EC2, a AWS oferece código do lado do cliente de exemplo para recuperar segredos do Secrets Manager em várias linguagens de programação conhecidas.
4. Revise periodicamente sua base de código e verifique novamente para confirmar se não há novos segredos adicionados ao código.
   - Considere usar uma ferramenta como o git-secrets para impedir a confirmação de novos segredos em seu repositório de código-fonte.
5. Monitore a atividade do Secrets Manager quanto a indicações de uso inesperado, acesso inadequado a segredos ou tentativas de excluir segredos.
6. Reduza a exposição humana às credenciais. Restrinja o acesso a credenciais de leitura, gravação e modificação a um perfil do IAM dedicado a esse fim, e apenas forneça acesso para assumir o perfil a um pequeno subconjunto de usuários operacionais.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciais temporárias
- SEC02-BP05 Fazer a auditoria e a alternância periódica das credenciais

Documentos relacionados:

- Conceitos básicos do AWS Secrets Manager
- Provedores de identidades e federação
- Amazon CodeGuru apresenta o Secrets Detector
- Como o AWS Secrets Manager usa o AWS Key Management Service
- Criptografia e descriptografia de segredos no Secrets Manager
- Entradas do blog do Secrets Manager
- Amazon RDS anuncia integração com o AWS Secrets Manager

Vídeos relacionados:

- Práticas recomendadas para gerenciar, recuperar e alternar segredos em grande escala
- Encontre segredos codificados com o Amazon CodeGuru Secrets Detector
- Proteger segredos para workloads híbridas usando o AWS Secrets Manager

Workshops relacionados:

- Armazenar, recuperar e gerenciar credenciais sigilosas no AWS Secrets Manager
- Ativações híbridas do AWS Systems Manager

SEC02-BP04 Contar com um provedor de identidades centralizado

Para identidades da força de trabalho (funcionários e prestadores de serviços), confie em um provedor de identidade que permita gerenciar identidades em um local centralizado. Isso facilita o gerenciamento do acesso em várias aplicações e sistemas, pois você está criando, atribuindo, gerenciando, revogando e auditando o acesso de um único local.

Resultado desejado: Você tem um provedor de identidade centralizado no qual gerencia centralmente os usuários da força de trabalho, as políticas de autenticação (como a exigência de autenticação multifator (MFA)) e a autorização para sistemas e aplicações (como atribuir acesso com base na associação ou nos atributos do grupo de um usuário). Os usuários da sua força de trabalho fazem login no provedor de identidade central e se federam (autenticação única) a aplicações internas e externas, eliminando a necessidade de os usuários se lembrarem de várias credenciais. Seu provedor de identidade é integrado aos seus sistemas de recursos humanos (RH) para que as mudanças de pessoal sejam automaticamente sincronizadas com seu provedor de identidade. Por exemplo, se alguém deixar sua organização, você poderá revogar automaticamente o acesso a aplicações e sistemas federados (inclusive a AWS). Você habilitou o registro em log detalhado de auditoria em seu provedor de identidade e está monitorando esses logs em busca de comportamentos incomuns do usuário.

Antipadrões comuns:

- Você não usa federação e autenticação única. Os usuários da sua força de trabalho criam contas de usuário e credenciais separadas em várias aplicações e sistemas.
- Você não automatizou o ciclo de vida das identidades dos usuários da força de trabalho, por exemplo, integrando seu provedor de identidade aos seus sistemas de RH. Quando um usuário deixa sua organização ou muda de função, você segue um processo manual para excluir ou atualizar seus registros em várias aplicações e sistemas.

Benefícios de estabelecer esta prática recomendada: Ao usar um provedor de identidade centralizado, você tem um único local para gerenciar as identidades e políticas dos usuários da força de trabalho, a capacidade de atribuir acesso às aplicações a usuários e grupos e a capacidade de monitorar a atividade de login do usuário. Ao se integrar aos seus sistemas de recursos humanos (RH), quando um usuário muda de função, essas alterações são sincronizadas com o provedor de identidade e atualizam automaticamente as aplicações e permissões atribuídas. Quando um usuário sai da sua organização, sua identidade é automaticamente desativada no provedor de identidade, revogando seu acesso a aplicações e sistemas federados.

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

Orientação para implementação

Orientação para usuários da força de trabalho que acessam a AWS

Usuários da força de trabalho em sua organização, como funcionários e prestadores de serviços, podem precisar acessar a AWS usando o AWS Management Console ou a AWS Command Line

Interface (AWS CLI) para realizar suas funções de trabalho. Você pode conceder acesso à AWS aos usuários da sua força de trabalho federando a partir de seu provedor de identidade centralizado para a AWS em dois níveis: federação direta para cada Conta da AWS ou federação para várias contas na organização da AWS.

- Para federar os usuários da sua força de trabalho diretamente com cada Conta da AWS, você pode usar um provedor de identidade centralizado para federar o AWS Identity and Access Management nessa conta. A flexibilidade do IAM permite que você habilite um SAML 2.0 ou um provedor de identidade Open ID Connect (OIDC) para cada Conta da AWS e use atributos de usuário federados para controle de acesso. Os usuários da sua força de trabalho usarão o navegador da web para fazer login no provedor de identidade fornecendo suas respectivas credenciais (como senhas e códigos de token MFA). O provedor de identidade emite uma declaração SAML para o navegador, que é enviada ao URL de login do AWS Management Console para permitir que o usuário faça autenticação única no AWS Management Console assumindo uma função do IAM. Seus usuários também podem obter credenciais temporárias de API da AWS para uso na AWS CLI ou em AWS SDKs pelo AWS STS assumindo a função do IAM usando uma declaração SAML do provedor de identidade.
- Para federar seus usuários da força de trabalho com várias contas em sua organização da AWS, você pode usar o AWS IAM Identity Center para gerenciar centralmente o acesso dos usuários de sua força de trabalho a Contas da AWS e aplicações. Você ativa o Centro de Identidade para sua organização e configura sua fonte de identidade. O IAM Identity Center fornece um diretório de origem de identidade padrão que você pode usar para gerenciar seus usuários e grupos. Como alternativa, você pode escolher uma fonte de identidade externa conectando-se ao seu provedor de identidade externo usando SAML 2.0 e provisionando automaticamente usuários e grupos usando o SCIM ou conectando-se ao diretório do Microsoft AD com o uso do AWS Directory Service. Depois que uma fonte de identidade é configurada, você pode atribuir acesso a usuários e grupos a Contas da AWS definindo políticas de privilégios mínimos em seus conjuntos de permissões. Os usuários da sua força de trabalho podem se autenticar por meio de seu provedor de identidade central para entrar no portal de acesso da AWS e autenticação única em Contas da AWS e aplicações em nuvem atribuídas a eles. Seus usuários podem configurar a AWS CLI v2 para se autenticar com o Centro de Identidade e obter credenciais para executar comandos da AWS CLI. O Centro de Identidade também permite acesso com autenticação única a aplicações da AWS, como o Amazon SageMaker Studio e portais do AWS IoT Sitewise Monitor.

Depois de seguir as orientações anteriores, os usuários da sua força de trabalho não precisarão mais usar IAM users e grupos para operações normais ao gerenciar workloads na AWS. Em vez disso,

seus usuários e grupos são gerenciados fora da AWS e os usuários podem acessar os recursos da AWS como uma identidade federada. As identidades federadas usam os grupos definidos pelo seu provedor de identidade centralizado. Você deve identificar e remover grupos do IAM, IAM users e credenciais de usuário de longa duração (senhas e chaves de acesso) que não são mais necessárias nas suas Contas da AWS. Você pode encontrar credenciais não utilizadas com o uso do relatórios de credenciais do IAM, excluindo IAM users correspondentes e excluindo grupos do IAM. Você pode aplicar uma política de controle de serviços (SCP) na sua organização, o que ajudará a impedir a criação de novos grupos e IAM users, forçando que esse acesso ocorra por meio de identidades federadas da AWS.

Orientação para usuários de suas aplicações

Você pode gerenciar as identidades dos usuários de suas aplicações, como um aplicativo móvel, usando o Amazon Cognito como seu provedor de identidade centralizado. O Amazon Cognito permite autenticação, autorização e gerenciamento de usuários de seus aplicativos móveis e da web. O Amazon Cognito fornece um repositório de identidades que pode ser expandido para milhões de usuários, oferece suporte à federação de identidades sociais e corporativas e oferece recursos avançados de segurança para ajudar a proteger seus usuários e sua empresa. Você pode integrar seu aplicativo web ou móvel personalizado ao Amazon Cognito para adicionar autenticação de usuário e controle de acesso aos seus aplicativos em minutos. Desenvolvido com base em padrões de identidade abertos, como SAML e Open ID Connect (OIDC), o Amazon Cognito oferece suporte a vários regulamentos de conformidade e se integra aos recursos de desenvolvimento de front-end e back-end.

Etapas da implementação

Etapas para usuários da força de trabalho acessarem a AWS

- Federe os usuários da sua força de trabalho à AWS usando um provedor de identidade centralizado de acordo com uma das seguintes abordagens:
  - Use o IAM Identity Center para habilitar a autenticação única para várias Contas da AWS em sua organização da AWS, federando com seu provedor de identidade.
  - Use o IAM para conectar seu provedor de identidade diretamente a cada Conta da AWS, permitindo acesso federado refinado.
- Identifique e remova IAM users e grupos que são substituídos por identidades federadas.

Etapas para usuários de suas aplicações

- Use o Amazon Cognito como um provedor de identidade centralizado para suas aplicações.
- Integre suas aplicações personalizadas com o Amazon Cognito usando o OpenID Connect e o OAuth. Você pode desenvolver suas aplicações personalizadas usando as bibliotecas do Amplify que fornecem interfaces simples para integração com uma variedade de serviços da AWS, como o Amazon Cognito para autenticação.

Recursos

Práticas recomendadas relacionadas ao Well-Architected:

- SEC02-BP06 Utilizar grupos e atributos de usuários
- SEC03-BP02 Conceder acesso com privilégio mínimo
- SEC03-BP06 Gerenciar o acesso com base no ciclo de vida

Documentos relacionados:

- Federação de identidades na AWS
- Práticas recomendadas de segurança no IAM
- Práticas recomendadas do AWS Identity and Access Management
- Introdução à administração delegada do IAM Identity Center
- Como usar políticas gerenciadas pelo cliente no IAM Identity Center para casos de uso avançados
- AWS CLI v2: provedor de credenciais do IAM Identity Center

Vídeos relacionados:

- AWS re:Inforce 2022 - AWS Identity and Access Management (IAM) deep dive (AWS re:Inforce 2022: aprofundamento no AWS Identity and Access Management (IAM))
- AWS re:Invent 2022 - Simplify your existing workforce access with IAM Identity Center (AWS re:Invent 2022: simplifique o acesso existente de sua força de trabalho com o IAM Identity Center)
- AWS re:Invent 2018: Mastering Identity at Every Layer of the Cake (AWS re:Invent 2018: dominar a identidade em todos os aspectos)

Exemplos relacionados:

- Workshop: Using AWS IAM Identity Center to achieve strong identity management (Uso do AWS IAM Identity Center para conseguir um forte gerenciamento de identidade)
- Workshop: Serverless identity (Identidade sem servidor)

Ferramentas relacionadas:

- Parceiros de competência em segurança da AWS: gerenciamento de identidade e acesso
- saml2aws

SEC02-BP05 Fazer a auditoria e a alternância periódica das credenciais

Audite e alterne as credenciais periodicamente para limitar o período durante o qual as credenciais podem ser usadas para acessar seus recursos. Credenciais de longo prazo criam muitos riscos, e estes podem ser reduzidos alternando credenciais de longo prazo regularmente.

Resultado desejado: implementar a alternância de credenciais para ajudar a reduzir os riscos associados ao uso de credenciais de longo prazo. Auditar e corrigir regularmente a não conformidade com políticas de alternância de credenciais.

Antipadrões comuns:

- Não auditar o uso de credenciais.
- Utilizar credenciais de longo prazo desnecessariamente.
- Utilizar credenciais de longo prazo e não alterná-las regularmente.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientação de implementação

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 substituam chaves de acesso por credenciais

temporárias. Ao migrar de usuários do AWS Identity and Access Management (IAM) para identidades centralizadas, é possível gerar um relatório de credenciais para fazer auditoria de seus usuários.

Também recomendamos implementar e monitorar a MFA no provedor de identidades. É possível configurar o Regras do AWS Config ou usar Padrões de segurança AWS Security Hub para monitorar se os usuários têm a MFA ativada. Considere utilizar o IAM Roles Anywhere para fornecer credenciais temporárias para identidades de máquina. Em situações em que o uso de perfis do IAM e credenciais temporárias não é possível, é necessário realizar auditoria frequente e alternar as chaves de acesso.

Etapas da implementação

- Fazer auditoria nas credenciais regularmente: a auditoria das identidades configuradas em seu provedor de identidades e no IAM ajuda a garantir que somente identidades autorizadas tenham acesso à sua workload. Essas identidades podem incluir, entre outros, usuários do IAM, do AWS IAM Identity Center, do Active Directory ou usuários em um provedor de identidades upstream diferente. Por exemplo, remova as pessoas que saem da organização e as funções entre contas que não são mais necessárias. Estabeleça um processo para auditar periodicamente as permissões para os serviços acessados por uma entidade do IAM. Isso ajuda a identificar as políticas que você precisa modificar a fim de remover todas as permissões não utilizadas. Use relatórios de credenciais e o AWS Identity and Access Management Access Analyzer para auditar credenciais e permissões do IAM. É possível utilizar o Amazon CloudWatch para configurar alarmes para chamadas de API específicas chamadas em seu ambiente da AWS. O Amazon GuardDuty também pode alertar você sobre atividade inesperada, que pode indicar acesso excessivamente permissivo ou acesso acidental às credenciais do IAM.
- Alternar credenciais regularmente: quando você não pode utilizar credenciais temporárias, alterne as chaves de acesso do IAM de longo prazo regularmente (no máximo, a cada 90 dias). Se uma chave de acesso for divulgada acidentalmente sem seu conhecimento, isso limitará o período de uso das credenciais para acessar seus recursos. Para ter informações sobre a alternância de chaves de acesso para usuários do IAM, consulte Alternar chaves de acesso.
- Revisar as permissões do IAM: para melhorar a segurança de sua Conta da AWS, revise e monitore regularmente cada uma das políticas do IAM. Verifique se as políticas seguem o princípio de privilégio mínimo.
- Considerar automatizar a criação e as atualizações dos recursos do IAM:o IAM Identity Center automatiza muitas tarefas do IAM, como o gerenciamento de perfis e políticas. Como alternativa, o AWS CloudFormation pode ser usado para automatizar a implantação de recursos do IAM, como

perfis e políticas, para reduzir a chance de erros humanos, pois os modelos podem ser verificados e ter controle de versão.

- Utilizar o IAM Roles Anywhere para substituir os usuários do IAM para identidades de máquina: o IAM Roles Anywhere possibilita usar perfis em áreas onde não seria possível tradicionalmente, como em servidores on-premises. O IAM Roles Anywhere utiliza um certificado X.509 confiável para realizar a autenticação na AWS e receber credenciais temporárias. O uso do IAM Roles Anywhere evita a necessidade de alternar essas credenciais, pois credenciais de longo prazo não são mais armazenadas em seu ambiente on-premises. Você precisará monitorar e alternar o certificado X.509 ao aproximar-se da validade.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciais temporárias
- SEC02-BP03 Armazenar e usar segredos com segurança

Documentos relacionados:

- Conceitos básicos do AWS Secrets Manager
- Práticas recomendadas do IAM
- Provedores de identidades e federação
- Soluções de parceiros de segurança: acesso e controle de acesso
- Credenciais de segurança temporárias
- Obter relatórios de credenciais da sua Conta da AWS

Vídeos relacionados:

- Práticas recomendadas para gerenciar, recuperar e alternar segredos em grande escala
- Gerenciar permissões de usuário em grande escala com o AWS IAM Identity Center
- Dominar a identidade em todos os aspectos

Exemplos relacionados:

- Well-Architected Lab: Limpeza automatizada de usuários do IAM

- Well-Architected Lab: Implantação automatizada de grupos e perfis do IAM

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

SEGURANÇA 3. Como gerenciar 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 em sua organização
- SEC03-BP09 Compartilhar recursos com segurança com terceiros

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

É prática recomendada conceder somente o acesso de que as identidades precisam para realizar ações em recursos específicos e sob condições específicas. Use grupos e atributos de identidade para definir permissões dinamicamente em 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, se um desenvolvedor sair do projeto, o acesso dele é automaticamente revogado sem alterar as políticas de acesso adjacentes.

Resultado desejado: os usuário somente têm permissões necessárias para fazerem seus respectivos trabalhos. Os usuários devem ter acesso apenas a ambientes de produção para realizar uma tarefa específica dentro de um período limitado e o acesso deve ser revogado quando a tarefa for concluída. As permissões devem ser revogadas quando não forem mais necessárias, incluindo quando um usuário for para um projeto diferente ou mudar de cargo. Privilégios de administrador devem ser concedidos apenas a um grupo pequeno de administradores confiáveis. As permissões devem ser revistas regularmente para evitar desvios de permissão. Contas de máquina ou sistema devem ter apenas o mínimo de permissões necessárias para concluir as tarefas.

Antipadrões comuns:

- Usar como padrão a concessão de permissões de administrador aos usuários.
- Usar o usuário raiz para atividades diárias.
- Criar políticas permissivas demais, mas sem privilégios completos de administrador.
- Não revisar as permissões para entender se elas permitem o acesso de privilégio mínimo.

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

Orientações para a implementação

O princípio de estados privilégio mínimo para as identidades deve ser apenas permitido para realizar o mínimo de ações necessárias para cumprir uma tarefa específica. Isso equilibra a usabilidade, eficiência e segurança. Operar sobre esse princípio ajuda a limitar acesso não intencional e a rastrear quem tem acesso a quais recursos. Usuários e perfis do IAM não têm permissões por padrão. O usuário raiz tem acesso total e deve ser controlado e monitorado rigidamente, além de usado apenas para tarefas que necessitam acesso raiz.

Políticas do IAM são usadas para conceder explicitamente permissões aos perfis do IAM ou recursos específicos. Por exemplo, políticas com base em identidade podem ser anexadas a grupos do IAM, enquanto buckets do S3 podem ser controlados por políticas baseadas em recursos.

Ao criar e associar uma política do IAM, você pode especificar as ações de serviço, os recursos e as condições que devem ser verdadeiras para que a AWS permita ou negue o acesso. A AWS oferece suporte a uma variedade de condições para ajudar você a reduzir o acesso. Por exemplo, ao usar `PrincipalOrgID` como chave de condição, você pode negar ações se o solicitante não for parte da sua organização da AWS.

Você também pode controlar as solicitações feitas pelos serviços da AWS em seu nome, como a criação, pelo AWS CloudFormation, de uma função do AWS Lambda, usando a chave de condição `CalledVia`. Tipos diferentes de política devem estar em camadas para estabelecer a defesa em profundidade e limitar as permissões gerais de seus usuários. Você pode restringir as permissões que podem ser concedidas e sob quais condições. Por exemplo, você pode permitir que suas equipes de aplicação criem suas próprias políticas do IAM para os sistemas que criam, mas deve também aplicar uma Fronteira de permissão para limitar o máximo de permissões que o sistema pode receber.

Etapas da implementação

- Implementar políticas de privilégio mínimo: atribua políticas de acesso com privilégio mínimo a grupos e perfis do IAM para refletir a função ou o perfil do usuário que você definiu.
  - Basear as políticas no uso da API: uma maneira de determinar as permissões necessárias é analisar os logs do AWS CloudTrail. Essa análise permite que você crie permissões personalizadas para as ações do usuário dentro da AWS. O IAM Access Analyzer pode gerar automaticamente uma IAM política com base na atividade. Você pode usar o IAM Access Advisor no nível da organização ou da conta para rastrear as últimas informações acessadas para determinada política.

- Considerar o uso de políticas gerenciadas da AWS para cargos. Pode ser difícil saber por onde começar ao criar políticas de permissões mais estritas. A AWS gerencia políticas para cargos comuns, como faturamento, administradores de banco de dados e cientistas de dados. Essas políticas podem ajudar a diminuir o acesso dos usuários ao determinar como implementar as políticas de privilégio mínimo.
- Remover permissões desnecessárias: remova permissões que não são necessárias e ajuste políticas muito permissivas. A geração de política pelo IAM Access Analyzer pode ajudar a ajustar as políticas de permissão.
- Garantir que os usuários tenham acesso limitado a ambientes de produção: os usuários devem ter acesso a ambientes de produção apenas com um caso de uso válido. Depois de o usuário realizar as tarefas específicas para as quais foi necessário o acesso à produção, o acesso deve ser revogado. Limitar o acesso a ambientes de produção evita eventos não intencionais e que causam impacto à produção, além de diminuir o escopo do impacto do acesso não intencional.
- Considerar os limites de permissões: um limite de permissões é um recurso 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.
- Considerar tags de recursos para permissões: um modelo de controle de acesso baseado em atributo que usa tags de recursos permite conceder acesso com base no propósito do recurso, proprietário, ambiente ou outros critérios. Por exemplo, você pode usar tags de recurso para diferenciar entre ambientes de desenvolvimento e produção. Ao usar essas tags, é possível restringir os desenvolvedores ao ambiente de desenvolvimento. Ao combinar as tags e as políticas de permissões, você consegue alcançar um acesso restrito ao recurso sem precisar definir políticas complicadas e personalizadas para cada cargo.
- Use políticas de controle de serviço para AWS Organizations. As políticas de controle de serviço controlam centralmente o máximo de permissões disponíveis para contas de membros em sua organização. É importante notar que as políticas de controle de serviço permitem que você restrinja as permissões do usuário raiz nas contas de membros. Considere também o uso do AWS Control Tower, que fornece controles gerenciados prescritivos que enriquecem o AWS Organizations. Também é possível definir os seus próprios controles no Control Tower.
- Estabelecer uma política de ciclo de vida para sua organização: as políticas de ciclo de vida do usuário definem tarefas a serem realizadas quando os usuários entram na AWS, mudam de cargo ou escopo de trabalho ou não precisam mais de acesso à AWS. As análises de permissão devem

ser feitas durante todas as etapas do ciclo de vida do usuário para verificar se as permissões estão adequadamente restritas e para evitar desvios nas permissões.

- Estabelecer uma programação regular para rever as permissões e remover as permissões desnecessárias: frequentemente, você deve verificar o acesso do usuário para garantir que ele não tenha acesso muito permissivo. O AWS Config e o IAM Access Analyzer podem ajudar ao auditar as permissões do usuário.
- Estabelecer uma matriz de cargos: uma matriz de cargos exibe os diversos cargos e níveis de acesso necessários dentro de sua área da AWS. Com uma matriz de cargos, você pode definir e separar as permissões com base nas responsabilidades do usuário dentro da sua organização. Use grupos em vez de aplicar permissões diretamente a usuários ou cargos individuais.

Recursos

Documentos relacionados:

- Conceder privilégio mínimo
- 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)
- Delegate permission management to developers by using IAM permissions boundaries (Delegar gerenciamento de permissões para desenvolvedores usando os limites de permissões do IAM)
- 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)

- Controle de acesso baseado em atributos (ABAC)
- Redução do escopo da política ao exibir a atividade do usuário
- Visualizar acesso do cargo
- Use a marcação para organizar seu ambiente e gerar responsabilidade
- Estratégias de marcação da AWS
- Marcação de recursos da AWS

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 users and roles to prevent privilege escalation? (Como posso usar limites de permissões para limitar usuários e funções e evitar escalação do privilégio?)

Exemplos relacionados:

- Lab: IAM permissions boundaries delegating role creation (Laboratório: limites de permissões do IAM que delegam a criação de perfis)
- Lab: IAM tag based access control for EC2 (Laboratório: controle de acesso baseado em tags do IAM para EC2)

SEC03-BP03 Estabelecer processo de acesso de emergência

Crie um processo que permita acesso emergencial às suas workloads no caso improvável de um problema com seu provedor de identidades centralizado.

Você deve criar processos para diferentes modos de falha que possam resultar em um evento de emergência. Por exemplo, em circunstâncias normais, os usuários da sua força de trabalho são federados na nuvem usando um provedor de identidades centralizado (SEC02-BP04) para gerenciar as respectivas workloads. No entanto, se o provedor de identidades centralizado falhar ou a configuração da federação na nuvem for modificada, talvez os usuários de sua força de trabalho não consigam se federar na nuvem. Um processo de acesso de emergência permite que administradores autorizados acessem seus recursos de nuvem por meios alternativos (como uma forma alternativa de federação ou acesso direto do usuário) para corrigir problemas com sua

configuração de federação ou workloads. O processo de acesso de emergência é usado até que o mecanismo normal de federação seja restaurado.

Resultado desejado:

- Você definiu e documentou os modos de falha que são considerados uma emergência: considere suas circunstâncias normais e os sistemas dos quais seus usuários dependem para gerenciar suas workloads. Pense em como cada uma dessas dependências pode falhar e causar uma situação de emergência. Você pode encontrar as perguntas e as práticas recomendadas no Pilar Confiabilidade útil para identificar modos de falha e arquitetar sistemas mais resilientes com o objetivo de minimizar a probabilidade de falhas.
- Você documentou as etapas que devem ser seguidas para confirmar uma falha como emergência. Por exemplo, é possível exigir que os administradores de identidade confiram o status de seus provedores de identidade primário e de reserva e, se nenhum dos dois estiver disponível, declarar um evento de emergência por falha do provedor de identidades.
- Você definiu um processo de acesso de emergência específico de cada tipo de modo de emergência ou falha. Ser específico pode reduzir a tentação de seus usuários de abusar de um processo geral para todos os tipos de emergência. Seus processos de acesso de emergência descrevem as circunstâncias em que cada processo deve ser usado e, inversamente, as situações em que o processo não deve ser usado e apontam para processos alternativos que podem ser aplicados.
- Seus processos são bem documentados com instruções detalhadas e manuais que podem ser seguidos com rapidez e eficiência. Lembre-se de que um evento de emergência pode ser um momento estressante para os usuários e eles podem estar sob extrema pressão de tempo, portanto, projete o processo para ser o mais simples possível.

Antipadrões comuns:

- Você não tem processos de acesso de emergência bem documentados e bem testados. Os usuários não estão preparados para uma emergência e seguem processos improvisados quando surge um evento de emergência.
- Seus processos de acesso de emergência dependem dos mesmos sistemas (como um provedor de identidades centralizado) que seus mecanismos de acesso normais. Isso significa que a falha desse sistema pode afetar os mecanismos de acesso normal e de emergência e prejudicar sua capacidade de se recuperar da falha.

- Seus processos de acesso de emergência são usados em situações não emergenciais. Por exemplo, os usuários frequentemente usam de forma indevida os processos de acesso de emergência, pois acham mais fácil fazer alterações diretamente do que enviá-las por meio de um pipeline.
- Seus processos de acesso de emergência não geram logs suficientes para auditar os processos, ou os logs não são monitorados para alertar sobre o possível uso indevido dos processos.

Benefícios de estabelecer esta prática recomendada:

- Com processos de acesso de emergência bem documentados e testados, é possível reduzir o tempo gasto pelos usuários para responder e resolver um evento de emergência. Isso pode resultar em menos tempo de inatividade e maior disponibilidade dos serviços fornecidos aos seus clientes.
- Você pode rastrear cada solicitação de acesso de emergência e detectar e alertar sobre tentativas não autorizadas de uso indevido do processo para eventos não emergenciais.

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

Orientação para implementação

Esta seção fornece orientação para criar processos de acesso de emergência para vários modos de falha relacionados às workloads implantadas na AWS, começando com uma orientação comum que se aplica a todos os modos de falha e seguida por uma orientação específica com base no tipo de modo de falha.

Orientação comum para todos os modos de falha

Pense no seguinte ao projetar um processo de acesso de emergência para um modo de falha:

- Documente as pré-condições e as suposições do processo: quando o processo deve ou não ser usado. Isso ajuda a detalhar o modo de falha e documentar suposições, como o estado de outros sistemas relacionados. Por exemplo, o processo do Modo de falha 2 pressupõe que o provedor de identidades está disponível, mas a configuração na AWS foi modificada ou expirou.
- Pré-crie os recursos necessários para o processo de acesso de emergência (SEC10-BP05). Por exemplo, crie previamente a Conta da AWS de acesso de emergência com IAM users e perfis e os perfis entre contas do IAM em todas as contas da workload. Isso verifica se esses recursos estão prontos e disponíveis quando ocorre um evento de emergência. Ao pré-criar recursos, você não depende das APIs do ambiente de gerenciamento da AWS (usadas para criar e modificar recursos

da AWS) que podem ficar indisponíveis em caso de emergência. Além disso, ao pré-criar recursos do IAM, você não precisa contabilizar possíveis atrasos devido à eventual consistência.

- Inclua processos de acesso de emergência como parte dos planos de gerenciamento de incidentes (SEC10-BP02). Documente como os eventos de emergência são acompanhados e comunicados a outras pessoas na organização, como equipes de colegas, sua liderança e, quando aplicável, externamente a seus clientes e parceiros de negócios.
- Defina o processo de solicitação de acesso de emergência no sistema de fluxo de trabalho de solicitação de serviço existente, caso haja um. Normalmente, esses sistemas de fluxo de trabalho permitem criar formulários de admissão para coletar informações sobre a solicitação, acompanhar a solicitação em cada estágio do fluxo de trabalho e adicionar etapas de aprovação automatizadas e manuais. Relacione cada solicitação a um evento de emergência correspondente acompanhado no sistema de gerenciamento de incidentes. Ter um sistema uniforme para acessos de emergência permite que você acompanhe essas solicitações em um único sistema, analise as tendências de uso e melhore os processos.
- Verifique se os processos de acesso de emergência só podem ser iniciados por usuários autorizados e exigem aprovações dos colegas ou da gerência do usuário, conforme apropriado. O processo de aprovação deve operar de forma eficaz dentro e fora do horário comercial. Defina como as solicitações de aprovação permitirão aprovadores secundários se os aprovadores primários não estiverem disponíveis e forem encaminhadas para a cadeia de gerenciamento até serem aprovadas.
- Verifique se o processo gera logs e eventos de auditoria detalhados para tentativas bem-sucedidas e fracassadas de obter acesso de emergência. Monitore o processo de solicitação e o mecanismo de acesso de emergência para detectar uso indevido ou acessos não autorizados. Correlacione a atividade com eventos de emergência contínuos do sistema de gerenciamento de incidentes e alerte quando as ações ocorrerem fora dos períodos esperados. Por exemplo, você deve monitorar e alertar sobre atividades na Conta da AWS de acesso de emergência, pois ela nunca deve ser usada em operações normais.
- Teste os processos de acesso de emergência periodicamente para verificar se as etapas estão claras e garantir o nível correto de acesso com rapidez e eficiência. Os processos de acesso de emergência devem ser testados como parte das simulações de resposta a incidentes (SEC10-BP07) e testes de recuperação de desastres (REL13-BP03).

Modo de falha 1: o provedor de identidades usado para federar na AWS não está disponível

Conforme descrito em SEC02-BP04 Contar com um provedor de identidades centralizado, recomendamos confiar em um provedor de identidades centralizado para federar os usuários de

sua força de trabalho e conceder acesso a Contas da AWS. Você pode federar em várias Contas da AWS na organização da AWS usando o IAM Identity Center ou federar em Contas da AWS individuais usando o IAM. Nos dois casos, os usuários da força de trabalho se autenticam com seu provedor de identidades centralizado antes de serem redirecionados a um endpoint de login da AWS para SSO.

No caso improvável do provedor de identidades centralizado não estar disponível, os usuários da sua força de trabalho não poderão se federar nas Contas da AWS nem gerenciar as workloads. Nesse evento de emergência, é possível fornecer um processo de acesso de emergência para um pequeno grupo de administradores acessar Contas da AWS a fim de realizar tarefas essenciais que não podem esperar até que seus provedores de identidades centralizados estejam online novamente. Por exemplo, seu provedor de identidades fica indisponível por quatro horas e, durante esse período, você precisa modificar os limites superiores de um grupo do Amazon EC2 Auto Scaling em uma conta de produção para lidar com um aumento inesperado no tráfego de clientes. Seus administradores de emergência devem seguir o processo de acesso de emergência a fim de obter acesso à Conta da AWS de produção específica e fazer as alterações necessárias.

O processo de acesso de emergência depende de uma Conta da AWS de acesso de emergência pré-criada usada exclusivamente para acesso de emergência e tem recursos da AWS (como perfis do IAM e IAM users) para apoiar o processo de acesso de emergência. Durante as operações normais, ninguém deve acessar a conta de acesso de emergência, e você deve monitorar e alertar sobre o uso indevido dessa conta (para receber mais detalhes, consulte a seção Orientação comum anterior).

A conta de acesso de emergência tem perfis do IAM de acesso de emergência com permissões para assumir perfis entre contas nas Contas da AWS que exigem acesso de emergência. Esses perfis do IAM são pré-criados e configurados com políticas de confiança que confiam nos perfis do IAM da conta de emergência.

O processo de acesso de emergência pode usar uma das seguintes abordagens:

- Você pode pré-criar um conjunto de IAM users para seus administradores de emergência na conta de acesso de emergência com senhas fortes e tokens de MFA associados. Esses IAM users têm permissões para assumir os perfis do IAM que permitem o acesso entre contas à Conta da AWS onde o acesso de emergência é necessário. Recomendamos criar o menor número possível de usuários e atribuir cada um a um único administrador de emergência. Durante uma emergência, um usuário administrador de emergência entra na conta de acesso de emergência usando sua senha e código de token MFA, muda para a função do IAM de acesso de emergência na conta de emergência e, por fim, muda para a função do IAM de acesso de emergência na conta da

workload para realizar a ação de alteração de emergência. A vantagem dessa abordagem é que cada IAM user é atribuído a um administrador de emergência, e você pode saber qual usuário fez login analisando os eventos do CloudTrail. A desvantagem é que você precisa manter vários IAM users com as respectivas senhas de longa duração e tokens de MFA associados.

- Você pode usar o usuário raiz da Conta da AWS de acesso de emergência para entrar na conta de acesso de emergência, assumir o perfil do IAM para acesso de emergência e assumir o perfil entre contas na conta da workload. Recomendamos definir uma senha forte e vários tokens de MFA para o usuário raiz. Também recomendamos armazenar a senha e os tokens de MFA em um cofre de credenciais corporativo seguro que imponha autenticação e autorização fortes. Você deve proteger a senha e os fatores de redefinição de tokens de MFA: defina o endereço de e-mail da conta como uma lista de distribuição de e-mail monitorada pelos administradores de segurança na nuvem e o número de telefone da conta como um número de telefone compartilhado que também seja monitorado pelos administradores de segurança. A vantagem dessa abordagem é que há um conjunto de credenciais de usuário raiz para gerenciar. A desvantagem é que, como se trata de um usuário compartilhado, vários administradores podem fazer login como usuário raiz. Você deve fazer auditoria dos eventos de log do cofre corporativo para identificar qual administrador fez check-out da senha do usuário raiz.

Modo de falha 2: a configuração do provedor de identidades na AWS foi modificada ou expirou

Para permitir que os usuários de sua força de trabalho sejam federados nas Contas da AWS, você pode configurar o IAM Identity Center com um provedor de identidades externo ou criar um provedor de identidades do IAM (SEC02-BP04). Normalmente, você os configura importando um documento XML de metadados SAML fornecido pelo provedor de identidades. O documento XML de metadados inclui um certificado X.509 correspondente a uma chave privada que o provedor de identidades usa para assinar as declarações SAML.

Essas configurações no lado da AWS podem ser modificadas ou excluídas por engano por um administrador. Em outro cenário, o certificado X.509 importado para a AWS pode expirar, e um novo XML de metadados com um novo certificado ainda não foi importado para a AWS. Os dois cenários podem interromper a federação na AWS para os usuários de sua força de trabalho, ocasionando uma emergência.

Nesse evento de emergência, você pode fornecer aos seus administradores de identidade acesso à AWS para resolver os problemas de federação. Por exemplo, seu administrador de identidade usa o processo de acesso de emergência para fazer login na Conta da AWS de acesso de emergência, muda para um perfil na conta de administrador do Centro de Identidade e atualiza a configuração do

provedor de identidades externo importando o documento XML de metadados SAML mais recente do provedor de identidades para reativar a federação. Depois que a federação for corrigida, os usuários da sua força de trabalho continuarão usando o processo operacional normal para federar em suas contas da workload.

Você pode seguir as abordagens detalhadas no Modo de falha 1 anterior para criar um processo de acesso de emergência. É possível conceder permissões de privilégio mínimo aos seus administradores de identidade a fim de acessar somente a conta de administrador do Centro de Identidade e realizar ações no Centro de Identidade nessa conta.

Modo de falha 3: interrupção do Centro de Identidade

No caso improvável de uma interrupção do IAM Identity Center ou da Região da AWS, recomendamos definir uma configuração que possa ser usada para conceder acesso temporário ao AWS Management Console.

O processo de acesso de emergência usa a federação direta do provedor de identidades no IAM em uma conta de emergência. Para receber detalhes sobre as considerações sobre o processo e o design, consulte Configurar o acesso de emergência ao AWS Management Console.

Etapas da implementação

Etapas comuns para todos os modos de falha

- Crie uma Conta da AWS dedicado aos processos de acesso de emergência. Pré-crie os recursos do IAM necessários na conta, como perfis do IAM ou IAM users e, opcionalmente, provedores de identidades do IAM. Além disso, crie previamente perfis do IAM entre contas nas Contas da AWS da workload com relacionamentos de confiança com os perfis do IAM correspondentes na conta de acesso de emergência. Você pode usar o AWS CloudFormation StackSets com AWS Organizations para criar esses recursos nas contas de membros de sua organização.

- Crie políticas de controle de serviço do AWS Organizations (SCPS) para negar a exclusão e a modificação dos perfis do IAM entre contas nas Contas da AWS de membros.
- Ative o CloudTrail para a Conta da AWS de acesso de emergência e envie os eventos da trilha a um bucket central do S3 em sua Conta da AWS de coleção de logs. Se você estiver usando o AWS Control Tower para configurar e controlar seu ambiente de várias contas da AWS, todas as contas que você criar usando o AWS Control Tower ou inscrever no AWS Control Tower terão o CloudTrail ativado por padrão e serão enviadas a um bucket do S3 em uma Conta da AWS de arquivo de log dedicado.

- Monitore a atividade da conta de acesso de emergência criando regras do EventBridge que correspondam ao login do console e à atividade da API pelos perfis de emergência do IAM. Envie notificações ao seu centro de operações de segurança quando ocorrerem atividades fora de um evento de emergência contínuo acompanhado no sistema de gerenciamento de incidentes.

Etapas adicionais para o Modo de falha 1: o provedor de identidades usado para federar na AWS não está disponível, e o Modo de falha 2: a configuração do provedor de identidades na AWS foi modificada ou expirou

- Pré-crie recursos de acordo com o mecanismo escolhido para acesso de emergência:
  - Usar IAM users: pré-crie-os IAM users com senhas fortes e dispositivos de MFA associados.
  - Usar o usuário raiz da conta de emergência: configure o usuário raiz com uma senha forte e armazene a senha no seu cofre de credenciais corporativo. Associe vários dispositivos físicos de MFA ao usuário raiz e armazene os dispositivos em locais que possam ser acessados rapidamente pelos membros de sua equipe de administradores de emergência.

Etapas adicionais para o Modo de falha 3: interrupção do Centro de Identidade

- Conforme detalhado em Configurar o acesso de emergência ao AWS Management Console, na Conta da AWS de acesso de emergência, crie um provedor de identidades do IAM para ativar a federação direta de SAML a partir do provedor de identidades.
- Crie grupos de operações de emergência no IdP sem membros.
- Crie perfis do IAM correspondentes aos grupos de operações de emergência na conta de acesso de emergência.

Recursos

Práticas recomendadas relacionadas ao Well-Architected:

- SEC02-BP04 Contar com um provedor de identidades centralizado
- SEC03-BP02 Conceder acesso com privilégio mínimo
- SEC10-BP02 Desenvolver planos de gerenciamento de incidentes
- SEC10-BP07 Promover dias de jogo

Documentos relacionados:

- Configurar o acesso de emergência ao AWS Management Console
- Permitir que usuários federados do SAML 2.0 acessem o AWS Management Console
- Acesso de emergência

Vídeos relacionados:

- AWS re:Invent 2022 - Simplify your existing workforce access with IAM Identity Center
- AWS re:Inforce 2022 - AWS Identity and Access Management (IAM) deep dive

Exemplos relacionados:

- Perfil de acesso de emergência da AWS
- AWS customer playbook framework
- AWS incident response playbook samples

SEC03-BP04 Reduzir as permissões continuamente

À medida que suas equipes determinarem o acesso de que precisam, remova as permissões desnecessárias e estabeleça processos de análise para obter permissões de privilégio mínimo. Monitore e remova continuamente identidades e permissões não utilizadas para acesso humano e de máquina.

Resultado desejado: as políticas de permissão devem seguir o princípio de privilégio mínimo. À medida que os cargos e os perfis se tornem mais bem definidos, suas políticas de permissões precisam ser analisadas para remover permissões desnecessárias. Essa abordagem reduz o escopo do impacto caso as credenciais sejam expostas de forma acidental ou sejam acessadas sem autorização.

Antipadrões comuns:

- Usar como padrão a concessão de permissões de administrador aos usuários.
- Criar políticas permissivas demais, mas sem privilégios completos de administrador.
- Manter as políticas de permissão quando não são mais necessárias.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientação de implementação

Enquanto as equipes e os projetos estiverem começando, políticas de permissão permissivas podem ser usadas para inspirar inovação e agilidade. Por exemplo, em um ambiente de desenvolvimento ou teste, os desenvolvedores podem receber acesso a uma ampla gama de serviços da AWS. Recomendamos avaliar o acesso de forma contínua e restringir o acesso somente àqueles serviços e ações de serviço necessários para concluir o trabalho atual. Recomendamos essa avaliação para identidades humanas e de máquina. Identidades de máquina, às vezes, denominadas contas de sistema ou serviço, são identidades que fornecem acesso da AWS a aplicações ou servidores. Esse acesso é especialmente importante em um ambiente de produção, em que as permissões excessivamente permissivas podem causar um grande impacto e expor dados dos clientes.

A AWS oferece vários métodos para ajudar a identificar usuários, perfis, permissões e credenciais não utilizados. A AWS também pode ajudar a analisar a atividade de acesso dos usuários e dos perfis do IAM, como chaves de acesso associadas, e o acesso aos recursos da AWS, como objetos em buckets do Amazon S3. A geração de políticas do AWS Identity and Access Management Access Analyzer pode auxiliar você a criar políticas de permissão restritivas com base nos serviços e nas ações reais com os quais uma entidade principal interage. O controle de acesso baseado em atributo (ABAC) pode ajudar a simplificar o gerenciamento de permissões, pois você pode conceder permissões aos usuários utilizando os atributos deles em vez de anexar políticas de permissões diretamente a cada usuário.

Etapas da implementação

- Utilizar o AWS Identity and Access Management Access Analyzer: o IAM Access Analyzer ajuda a identificar os recursos na organização e nas contas, como buckets do Amazon Simple Storage Service (Amazon S3) ou perfis do IAM, que são compartilhados com uma entidade externa.
- Utilizar a geração de políticas do IAM Access Analyzer: a geração de políticas do IAM Access Analyzer ajuda você a criar políticas de permissão detalhadas com base em um usuário do IAM ou na atividade de acesso de um perfil.
- Determinar um cronograma e uma política de uso aceitáveis para usuários e perfis do IAM: utilize o carimbo de data e hora de último acesso para identificar usuários e perfis não utilizados e removê-los. Revise as informações de serviço e ação acessadas mais recentemente para identificar e definir o escopo das permissões para usuários e perfis específicos. Por exemplo, você pode usar as informações acessadas mais recentemente para identificar as ações específicas do Amazon S3 exigidas pelo perfil da aplicação e restringir o acesso do perfil apenas a essas ações. Os recursos de informações acessadas mais recentemente 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.

- Considerar o registro em log dos eventos de dados no AWS CloudTrail: por padrão, o CloudTrail não registra eventos de dados, como atividade em nível de objeto do Amazon S3 (por exemplo, `GetObject` e `DeleteObject`) ou atividades de tabelas do Amazon DynamoDB (por exemplo, `PutItem` e `DeleteItem`). Considere ativar o registro em log desses eventos para determinar quais usuários e perfis precisam acessar objetos do Amazon S3 ou itens de tabelas do DynamoDB específicos.

Recursos

Documentos relacionados:

- Conceder privilégio mínimo
- Remova credenciais desnecessárias
- O que é o AWS CloudTrail?
- Trabalhando com políticas
- Registrar em log e monitorar no DynamoDB
- Habilitar o log de eventos do CloudTrail para buckets e objetos do Amazon S3
- Obter relatórios de credenciais da sua Conta da AWS

Vídeos relacionados:

- Torne-se um mestre em políticas do IAM em 60 minutos ou menos
- Separação de tarefas, privilégio mínimo, delegação e CI/CD
- AWS re:Inforce 2022: Aprofundamento no AWS Identity and Access Management (IAM)

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 específicos que exigem esse acesso.

Resultado desejado: saber quais de seus recursos da AWS são compartilhados e com quem. Monitorar e auditar continuamente seus recursos compartilhados para verificar se eles são compartilhados com apenas entidades principais autorizadas.

Antipadrões comuns:

- Não manter um inventário dos recursos compartilhados.
- Não seguir um processo de aprovação do acesso público ou entre contas aos recursos.

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

Orientação de implementação

Se a sua conta estiver no AWS Organizations, você poderá conceder acesso aos recursos à toda a organização, a unidades organizacionais específicas ou a contas individuais. Se sua conta não for membro de uma organização, você poderá compartilhar recursos com contas individuais. Você pode

conceder acesso direto entre contas usando políticas baseadas em recursos, por exemplo, políticas de buckets do Amazon Simple Storage Service (Amazon S3) ou permitindo que uma identidade principal em outra conta assuma um perfil do IAM em sua conta. Ao utilizar políticas de recursos, verifique se o acesso é concedido apenas a entidades principais autorizadas. Defina um processo para aprovar todos os recursos que devem ser acessíveis publicamente.

O AWS Identity and Access Management Access Analyzer utiliza segurança demonstrável 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 configurar o IAM Access Analyzer com o AWS Organizations para verificar se você tem visibilidade a todas as suas contas. O IAM Access Analyzer também possibilita que você visualize descobertas antes de implantar permissões de recursos. Isso permite validar que as alterações de política concedam apenas o acesso público e entre contas pretendido aos seus recursos. Ao projetar o acesso a várias contas, é possível utilizar políticas de confiança para controlar em quais casos um perfil pode ser assumido. Por exemplo, você pode usar a chave de condição `PrincipalOrgId` para negar uma tentativa de assumir um perfil de fora de seu AWS Organizations.

O AWS Config pode relatar recursos configurados incorretamente, e por meio de verificações de politica do AWS Config, pode detectar recursos que tenham acesso público configurado. Serviços, como AWS Control Tower e AWS Security Hub simplificam a implantação de controles de detecção e barreiras de proteção nos 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 é restaurado por Contas da AWS.

Etapas da implementação

- Pensar em ativar o AWS Config para AWS Organizations: o AWS Config permite que você agregue as descobertas de várias contas em um AWS Organizations em uma conta de administrador delegada. Isso oferece uma visão abrangente e permite que você implante o Regras do AWS Config nas contas para detectar recursos acessíveis ao público.
- Configurar o AWS Identity and Access Management Access Analyzer o IAM Access Analyzer ajuda a identificar os recursos na organização e nas contas, como buckets do Amazon S3 ou perfis do IAM, que são compartilhados com uma entidade externa.
- Usar autocorreção no AWS Config para responder a alterações na configuração do acesso público de buckets do Amazon S3: é possível reativar automaticamente as configurações de acesso público de bloco para buckets do Amazon S3.

- Implementar o monitoramento e os alertas para identificar se os buckets do Amazon S3 se tornaram públicos: é necessário ter o monitoramento e os alertas implementados para identificar quando o acesso público de blocos do Amazon S3 foi desativado e se os buckets do Amazon S3 se tornaram públicos. Além disso, se você estiver usando o AWS Organizations, poderá criar uma política de controle de serviços que impeça alterações nas políticas de acesso público do Amazon S3. O AWS Trusted Advisor confere se há buckets do Amazon S3 com permissões de acesso abertas. As permissões de bucket que concedem, upload ou excluem acesso a todos criam possíveis problemas de segurança, pois permitem que qualquer pessoa adicione, modifique ou remova itens em um bucket. A verificação do Trusted Advisor examina as permissões de bucket explícitas e as políticas de bucket associadas que podem substituir as permissões de bucket. Você também pode utilizar o AWS Config para monitorar seus buckets do Amazon S3 para acesso público. Para ter mais informações, consulte Como usar o AWS Config para monitorar e responder a buckets do Amazon S3 que possibilitam acesso público. Ao revisar o acesso, é importante considerar quais tipos de dados estão contidos em buckets do Amazon S3. O Amazon Macie ajuda a descobrir e proteger dados sigilosos, como PII, PHI e credenciais, como chaves privadas ou da AWS.

Recursos

Documentos relacionados:

- Usar o AWS Identity and Access Management Access Analyzer
- Biblioteca de controles do AWS Control Tower
- Norma de práticas de segurança básicas da AWS
- Regras gerenciadas do AWS Config
- Referência de verificação do AWS Trusted Advisor
- Monitorar resultados da verificação do AWS Trusted Advisor com o Amazon EventBridge
- Gerenciar regras do AWS Config em todas as contas de sua organização
- AWS Config e AWS Organizations

Vídeos relacionados:

- Práticas recomendadas para proteger seu ambiente de várias contas
- Análise aprofundada do IAM Access Analyzer

SEC03-BP08 Compartilhar recursos com segurança em sua organização

À medida que o número de workloads aumenta, talvez você precise compartilhar o acesso aos recursos nessas workloads ou fornecer os recursos várias vezes nas contas. Você pode ter estruturas para fragmentar seu ambiente, como ter ambientes de desenvolvimento, teste e produção. No entanto, ter estruturas de separação não limita o compartilhamento seguro. Ao compartilhar componentes que se sobrepõem, você pode reduzir a sobrecarga operacional e possibilitar uma experiência consistente sem precisar adivinhar o que ignorou ao criar o mesmo recurso várias vezes.

Resultado desejado: minimizar o acesso acidental utilizando métodos seguros para compartilhar recursos com sua organização e ajudar com sua iniciativa de prevenção de perda de dados. Reduza sua sobrecarga operacional em comparação com o gerenciamento de componentes individuais, reduza os erros gerados pela criação manual do mesmo componente várias vezes e aumente a escalabilidade de suas workloads. É possível se beneficiar da redução de tempo para a resolução em cenários de falhas em vários pontos e aumentar sua confiança na determinação de quando um componente não é mais necessário. Para ter orientações prescritivas sobre como analisar recursos compartilhados externamente, consulte SEC03-BP07 Analisar o acesso público e entre contas.

Antipadrões comuns:

- Falta de um processo para monitorar de forma contínua e alertar automaticamente sobre o compartilhamento externo inesperado.
- Falta de referência sobre o que deve ou não ser compartilhado.
- Ter como padrão uma política amplamente aberta em vez de compartilhar explicitamente quando necessário.
- Criar manualmente recursos básicos que se sobrepõem quando necessário.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientação de implementação

Projete seus controles e padrões de acesso para reger o consumo de recursos compartilhados com segurança e somente com entidades confiáveis. Monitore recursos compartilhados e revise o acesso a eles de forma contínua e seja alertado sobre o compartilhamento inadequado ou inesperado. Leia Analisar o acesso público e entre contas para ajudar você a estabelecer a governança a fim de reduzir o acesso externo apenas aos recursos que precisem dele e estabelecer um processo para monitorar de forma contínua e alertar automaticamente.

O compartilhamento entre contas no AWS Organizations é compatível com uma série de serviços da AWS, como o AWS Security Hub, Amazon GuardDuty e o AWS Backup. Esses serviços possibilitam compartilhar os dados em uma conta central, acessá-los ou gerenciar recursos e dados dessa conta. Por exemplo, o AWS Security Hub pode transferir as descobertas de contas individuais para uma conta central onde é possível visualizar todas elas. O AWS Backup pode realizar um backup de um recurso e compartilhá-lo entre contas. É possível utilizar o AWS Resource Access Manager (AWS RAM) para compartilhar outros recursos comuns, como sub-redes de VPC e anexos do Transit Gateway, AWS Network Firewall ou pipelines Amazon SageMaker.

Para restringir sua conta para somente compartilhar recursos em sua organização, utilize políticas de controle de serviços (SCPs) para impedir o acesso a entidades principais externas. Ao compartilhar recursos, combine controles baseados em identidade e controles de rede para criar um perímetro de dados para sua organização a fim de ajudar a proteger contra o acesso acidental. Um perímetro de dados é um conjunto de barreiras de proteção preventivas que ajudam a garantir que apenas suas identidades confiáveis acessem recursos confiáveis das redes esperadas. Esses controles impõem limites apropriados 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 um perímetro de dados, é possível usar políticas de endpoint de VPC e a condição `AWS:PrincipalOrgId` para garantir que as identidades que acessam seus buckets do Amazon S3 pertençam à sua organização. É importante observar que as SCPs não se aplicam a perfis vinculados a serviço (LSR) nem a entidades principais de serviços da AWS.

Ao utilizar o Amazon S3, desative as ACLs de seu bucket do Amazon S3 e utilize políticas do IAM para definir o controle de acesso. Para restringir o acesso a uma origem do Amazon S3 a partir do Amazon CloudFront, migre da identidade do acesso de origem (OAI) para um controle de acesso de origem (OAC), que é compatível com recursos adicionais, por exemplo, a criptografia do lado do servidor com o AWS Key Management Service.

Em alguns casos, convém permitir o compartilhamento de recursos fora de sua organização ou conceder a terceiros acesso aos seus recursos. Para ter orientações prescritivas sobre o gerenciamento de permissões para compartilhar recursos externamente, consulte Gerenciamento de permissões.

Etapas da implementação

1. Utilize o AWS Organizations.

   O AWS Organizations é um serviço de gerenciamento de contas que permite consolidar várias Contas da AWS em uma organização que você cria e gerencia centralmente. É possível agrupar

suas contas em unidades organizacionais (UOs) e anexar políticas diferentes a cada UO a fim de ajudar a atender às suas necessidades orçamentárias, de segurança e conformidade. Também é possível controlar como serviços de inteligência artificial (IA) e machine learning (ML) da AWS podem coletar e armazenar dados e usar o gerenciamento de várias contas dos serviços da AWS integrados ao Organizations.

2. Integre o AWS Organizations aos serviços da AWS.

   Ao ativar um serviço da AWS para realizar tarefas em seu nome nas contas membros de sua organização, o AWS Organizations cria um perfil vinculado a serviço do IAM para esse serviço em cada conta membro. Você deve gerenciar o acesso confiável usando o AWS Management Console, as APIs da AWS ou a AWS CLI. Para ter orientações prescritivas sobre como ativar o acesso confiável, consulte Usar o AWS Organizations com outros serviços da AWS e Serviços da AWS que podem ser usados com o Organizations.

3. Estabeleça um perímetro de dados.

   O perímetro da AWS, geralmente, é representado como uma organização gerenciada pelo AWS Organizations. Junto com redes e sistemas on-premises, o acesso a recursos da AWS é o que muitos consideram o perímetro de My AWS. O objetivo do perímetro é garantir que o acesso seja permitido se a identidade e o recurso forem confiáveis e a rede for esperada.

   a. Defina e implante os perímetros.

      Siga as etapas descritas em Implementação do perímetro do whitepaper Criar um perímetro na AWS para cada condição de autorização. Para ter orientações prescritivas sobre como proteger a camada de rede, consulte Proteção de redes.

   b. Monitore e alerte de forma contínua.

      O AWS Identity and Access Management Access Analyzer ajuda a identificar os recursos na organização e nas contas que são compartilhados com entidades externas. É possível integrar o IAM Access Analyzer ao AWS Security Hub para enviar e agregar as descobertas para um recurso do IAM Access Analyzer para o Security Hub a fim de ajudar a analisar o procedimento de segurança de seu ambiente. Para ativar a integração, ative o IAM Access Analyzer e o Security Hub em cada região em cada conta. Também é possível utilizar o Regras do AWS Config para fazer auditoria da configuração e alertar a parte adequada utilizando o AWS Chatbot com o AWS Security Hub. Depois, você pode utilizar Documentos de automação do AWS Systems Manager para corrigir os recursos sem conformidade.

   c. Para ter orientações prescritivas sobre como monitorar e alertar de forma contínua sobre recursos compartilhados externamente, consulte Analisar o acesso público e entre contas.

4. Utilize o compartilhamento de recursos em serviços da AWS e restrinja-o adequadamente.

   Muitos serviços da AWS possibilitam compartilhar recursos com outra conta ou almejar um recurso em outra conta, como Imagens de máquina da Amazon (AMIs) e AWS Resource Access Manager (AWS RAM). Restrinja a API `ModifyImageAttribute` para especificar as contas confiáveis com as quais compartilhar a AMI. Especifique a condição `ram:RequestedAllowsExternalPrincipals` ao utilizar o AWS RAM para restringir o compartilhamento somente à sua organização, a fim de ajudar a impedir o acesso de identidades não confiáveis. Para ter orientações prescritivas e considerações, consulte Compartilhamento de recursos e destinos externos.

5. Utilize o AWS RAM para compartilhar com segurança em uma conta ou com outras Contas da AWS.

   O AWS RAM ajuda você a compartilhar com segurança os recursos criados com perfis e usuários em sua conta e com outras Contas da AWS. Em um ambiente de várias contas, o AWS RAM possibilita criar um recurso uma vez e compartilhá-lo com outras contas. Essa abordagem ajuda a reduzir sua sobrecarga operacional ao oferecer consistência, visibilidade e capacidade de auditoria por meio de integrações com o Amazon CloudWatch e o AWS CloudTrail, o que você não recebe ao utilizar o acesso entre contas.

   Se você tiver recursos compartilhados anteriormente com o uso de uma política baseada em recurso, é possível utilizar a API `PromoteResourceShareCreatedFromPolicy` ou equivalente a fim de promover o compartilhamento de recursos para um compartilhamento completo de recursos do AWS RAM.

   Em alguns casos, convém realizar etapas adicionais para compartilhar recursos. Por exemplo, para compartilhar um snapshot criptografado, é necessário compartilhar uma chave do AWS KMS.

Recursos

Práticas recomendadas relacionadas:

- SEC03-BP07 Analisar o acesso público e entre contas
- SEC03-BP09 Compartilhar recursos com segurança com terceiros
- SEC05-BP01 Criar camadas de rede

Documentos relacionados:

- Proprietário do bucket concede permissão entre contas a objetos que não possui
- Como usar políticas de confiança com o IAM
- Criar um perímetro de dados na AWS
- Como usar um ID externo ao conceder acesso aos seus recursos da AWS para terceiros
- Serviços da AWS que podem ser usados com o AWS Organizations
- Estabelecer um perímetro de dados na AWS: permitir apenas que identidades confiáveis acessem os dados da empresa

Vídeos relacionados:

- Acesso granular com o AWS Resource Access Manager
- Como proteger seu perímetro de dados com endpoints da VPC
- Estabelecer um perímetro de dados na AWS

Ferramentas relacionadas:

- Exemplos de política de perímetro de dados

SEC03-BP09 Compartilhar recursos com segurança com terceiros

A segurança de seu ambiente de nuvem não é interrompida em sua organização. Sua organização pode contar com terceiros para gerenciar uma parte de seus dados. O gerenciamento de permissões para o sistema gerenciado por terceiros deve seguir a prática de acesso just-in-time utilizando o princípio de privilégio mínimo com credenciais temporárias. A trabalhar em parceria com terceiros, é possível reduzir o escopo do impacto e o risco de acesso acidental.

Resultado desejado: credenciais do AWS Identity and Access Management (IAM) de longo prazo, chaves de acesso do IAM e chaves secretas associadas a um usuário podem ser usadas por qualquer pessoa desde que as credenciais sejam válidas e ativas. O uso de um perfil do IAM e credenciais temporárias ajuda você a melhorar seu procedimento de segurança geral reduzindo o esforço para manter credenciais de longo prazo, inclusive o gerenciamento e a sobrecarga operacional dessas informações sigilosas. Ao utilizar um identificador universalmente exclusivo (UUID) para o ID externo na política de confiança do IAM e manter as políticas do IAM anexadas ao perfil do IAM sob seu controle, é possível fazer auditoria e garantir que o acesso concedido a terceiros não seja permissivo demais. Para ter orientações prescritivas sobre como analisar recursos compartilhados externamente, consulte SEC03-BP07 Analisar o acesso público e entre contas.

Antipadrões comuns:

- Utilizar a política de confiança do IAM padrão sem condições.
- Utilizar credenciais e chaves de acesso de longo prazo do IAM.
- Reutilizar IDs externos.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientação de implementação

Talvez você deseje permitir o compartilhamento de recursos fora do AWS Organizations ou conceder a terceiros acesso à sua conta. Por exemplo, um parceiro (terceiros) pode oferecer uma solução de monitoramento que precise acessar recursos em sua conta. Nesses casos, crie um perfil entre contas do IAM somente com os privilégios necessários para o parceiro. Além disso, defina uma política de segurança com o uso da condição de ID externo. Ao utilizar um ID externo, você ou o parceiro pode gerar um ID exclusivo para cada cliente, terceiros ou locação. O ID exclusivo não deve ser controlado por ninguém, exceto por você, depois de criado. O parceiro deve implementar um processo para relacionar o ID externo ao cliente de forma segura, auditável e reproduzível.

Também é possível usar o IAM Roles Anywhere para gerenciar perfis do IAM para aplicações fora do AWS que utilizam APIs da AWS.

Se o parceiro não precisar mais de acesso ao seu ambiente, remova o perfil. Evite fornecer credenciais de longo prazo para terceiros. Esteja ciente de outros serviços da AWS compatíveis com o compartilhamento. Por exemplo, o AWS Well-Architected Tool possibilita o compartilhamento de uma workload com outras Contas da AWS, e o AWS Resource Access Manager ajuda você a compartilhar com segurança um recurso da AWS que você possua com outras contas.

Etapas da implementação

1. Utilize perfis entre contas para fornecer acesso a contas externas.

   Os perfis entre contas reduzem a quantidade de informações sigilosas armazenadas por contas externas e terceiros para atender aos clientes. Os perfis entre contas possibilitam a você conceder acesso a recursos da AWS em sua conta de forma segura a terceiros, como AWS Partners ou outras contas em sua organização e, ao mesmo tempo, manter a capacidade de gerenciar e auditar esse acesso.

O parceiro pode oferecer serviço a você a partir de uma infraestrutura híbrida ou, como alternativa, extrair dados de um local externo. O IAM Roles Anywhere ajuda você a possibilitar que workloads de terceiros interajam com segurança com suas workloads da AWS e reduzir ainda mais a necessidade de credenciais de longo prazo.

Você não deve usar credenciais ou chaves de acesso de longo prazo associadas a usuários para conceder acesso a contas externas. Em vez disso, utilize perfis entre contas para conceder acesso entre contas.

2. Utilize um ID externo com terceiros.

   O uso de um ID externo possibilita designar quem pode assumir um perfil em uma política de confiança do IAM. A política de confiança pode exigir que o usuário que assume o perfil imponha a condição e o destino no qual ele está operando. Ele também fornece uma maneira para que o proprietário da conta permita que a função seja assumida somente em circunstâncias específicas. A função principal do ID externo é resolver e evitar o problema de substituto confuso.

   Utilize um ID externo se você for proprietário de uma Conta da AWS e tiver configurado um perfil para terceiros que acesse outras Contas da AWS além da sua, ou quando você pode assumir perfis em nome de clientes diferentes. Trabalhe com terceiros ou a AWS Partner para estabelecer uma condição de ID externo a ser incluída na política de confiança do IAM.

3. Utilize IDs externos universalmente exclusivos.

   Implemente um processo que gere um valor exclusivo aleatório para um ID externo, como um identificador universalmente exclusivo (UUID). Um parceiro que reutilize IDs externos entre diferentes clientes não resolve o problema de substituto confuso porque o cliente A pode ser capaz de visualizar dados do cliente B utilizando o ARN do perfil do cliente B junto com o ID externo duplicado. Em um ambiente de vários locatários, em que um parceiro atende a vários clientes com diferentes Contas da AWS, o parceiro deve usar um ID exclusivo diferente como o ID externo de cada Conta da AWS. O parceiro é responsável por detectar IDs externos duplicados e mapear de forma segura cada cliente ao seu respectivo ID externo. O parceiro deve testar para verificar se ele pode assumir o perfil somente ao especificar o ID externo. O parceiro deve evitar armazenar o ARN do perfil do cliente e o ID externo até que este seja necessário.

   O ID externo não é tratado como segredo, mas ele não pode ser um valor facilmente dedutível, como um número de telefone, um nome ou o ID da conta. Torne o ID externo um campo somente leitura de forma que o ID externo não possa ser alterado com o fim de representar a configuração.

Você ou o parceiro podem gerar o ID externo. Defina um processo para determinar quem é responsável pela geração do ID. Seja qual for a entidade que crie o ID externo, o parceiro impõe a exclusividade e os formatos de forma consistente entre os clientes.

4. Deprecie credenciais de longo prazo fornecidas pelo cliente.

   Deprecie o uso de credenciais de longo prazo e use perfis entre clientes ou o IAM Roles Anywhere. Se você precisar utilizar credenciais de longo prazo, estabeleça um plano para migrar para um acesso baseado em perfil. Para obter detalhes sobre como gerenciar chaves, consulte Gerenciamento de identidades. Trabalhe também com a equipe de sua Conta da AWS e o parceiro para estabelecer um runbook de mitigação de riscos. Para ter orientações prescritivas sobre como responder e mitigar o impacto em potencial do incidente de segurança, consulte Resposta a incidentes.

5. Verifique se a configuração tem orientações prescritivas ou é automatizada.

   A política criada para acesso entre contas em suas contas deve seguir o princípio de privilégio mínimo. O parceiro deve fornecer um documento de política de perfil ou um mecanismo de configuração automatizada que utilize um modelo do AWS CloudFormation ou um equivalente para você. Isso reduz a chance de erros associados à criação manual de políticas e oferece uma trilha auditável. Para ter mais informações sobre como usar um modelo do AWS CloudFormation para criar perfis entre contas, consulte Perfis entre contas.

   O parceiro deve fornecer um mecanismo de configuração automatizado e auditável. No entanto, ao utilizar o documento de política de perfis que descreve o acesso necessário, você deve automatizar a configuração do perfil. Com um modelo do AWS CloudFormation ou equivalente, você deve monitorar alterações com detecção de desvios como parte da prática de auditoria.

6. Considere alterações.

   Sua estrutura de contas, sua necessidade de terceiros ou a oferta de serviço pode ser alterada. Você deve antecipar alterações e falhas e planejar adequadamente com as pessoas, o processo e a tecnologia corretos. Audite o nível de acesso que você concede periodicamente e implemente métodos de detecção para alertar você de alterações inesperadas. Monitore e audite o uso do perfil e o datastore dos IDs externos. Você deve estar preparado para revogar o acesso de terceiros, seja de forma temporária ou permanente, como resultado de alterações ou padrões de acesso inesperados. Além disso, meça o impacto de sua operação de revogação, inclusive o tempo para realizá-la, as pessoas envolvidas, o custo e o impacto de outros recursos.

Para ter orientações prescritivas sobre métodos de detecção, consulte as Práticas recomendadas de detecção.

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciais temporárias
- 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
- SEC04 Detecção

Documentos relacionados:

- Proprietário do bucket concede permissão entre contas a objetos que não possui
- Como usar políticas de confiança com os perfis do IAM
- Delegar acesso entre Contas da AWS usando funções do IAM
- Como acesso recursos em outra Conta da AWS usando o IAM?
- Práticas recomendadas de segurança no IAM
- Lógica de avaliação de política entre contas
- Como usar um ID externo ao conceder acesso a seus recursos da AWS a terceiros
- Coletar informações de recursos do AWS CloudFormation criados em contas externas com recursos personalizados
- Usar ID externo com segurança para acessar contas da AWS pertencentes a outros
- Estender perfis do IAM fora do IAM com IAM Roles Anywhere)

Vídeos relacionados:

- Como permito que usuários ou perfis em uma Conta da AWS separada acessem minha Conta da AWS?
- AWS re:Invent 2018: Torne-se um mestre em políticas do IAM em 60 minutos ou menos
- AWSPráticas recomendadas do IAM e decisões de design

Exemplos relacionados:

- Well-Architected Lab: Assumir perfil do IAM entre contas do Lambda (Nível 300)
- Configurar o acesso entre contas ao Amazon DynamoDB
- AWS STS Network Query Tool

Detecção

Pergunta

- SEGURANÇA 4. Como detectar e investigar eventos de segurança?

SEGURANÇA 4. Como detectar e investigar 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

Retenha logs de eventos de segurança de serviços e aplicações. Esse é um princípio fundamental de segurança para auditoria, investigações e casos de uso operacionais e um requisito de segurança comum orientado por padrões, políticas e procedimentos de governança, risco e conformidade (GRC).

Resultado desejado: uma organização deve ser capaz de recuperar de forma confiável e consistente logs de ventos de segurança de serviços e aplicações da AWS de modo pontual quando necessário a fim de cumprir um processo ou obrigação interna, como resposta a incidentes de segurança. Considere centralizar os logs para ter melhores resultados operacionais.

Antipadrões comuns:

- Os logs são armazenados de forma perpétua ou excluídos muito precocemente.
- Todos podem acessar os logs.
- Contar inteiramente com processos manuais para uso e governança de logs.
- Armazenar todos os tipos de log em caso de necessidade.
- Conferir a integridade dos logs apenas quando necessário.

Benefícios do estabelecimento desta prática recomendada: implementar um mecanismo de análise da causa raiz (RCA) para incidentes de segurança e uma fonte de evidências para suas obrigações de governança, risco e conformidade.

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

Orientação de implementação

Durante uma investigação de segurança ou outros casos de uso com base em seus requisitos, você precisa ser capaz de analisar os logs relevantes a fim de registrar e entender o escopo total e a linha do tempo do incidente. Os logs também são necessários para geração de alertas indicando que ocorreram determinadas ações de interesse. É essencial selecionar, ativar, armazenar e configurar mecanismos de consulta, recuperação e alertas.

Etapas da implementação

- Selecione e ative fontes de logs. Antes de uma investigação de segurança, você precisa capturar logs relevantes para reconstruir, de forma retroativa a atividade em uma Conta da AWS. Selecione e ative fontes de logs relevantes para suas workloads.

  Os critérios de seleção de fonte de logs devem se basear nos casos de uso necessários à sua empresa. Estabeleça uma trilha para cada Conta da AWS utilizando o AWS CloudTrail ou uma trilha de AWS Organizations e configure um bucket do Amazon S3 para ela.

  O AWS CloudTrail é um serviço de registro em log que rastreia chamadas de API feitas em uma Conta da AWS capturando a atividade do serviço da AWS. É ativado por padrão com uma retenção de 90 dias de eventos de gerenciamento que podem ser recuperados por meio do histórico de eventos do CloudTrail utilizando o AWS Management Console, a AWS CLI ou um AWS SDK. Para ter uma retenção maior e visibilidade dos eventos de dados, crie uma trilha do CloudTrail e associe-a a um bucket do Amazon S3 e opcionalmente com um grupo de logs do Amazon CloudWatch. Como alternativa, você pode criar um CloudTrail Lake, que retém logs do CloudTrail por até sete anos e oferece um recursos e consultas baseado em SQL

A AWS recomenda que os clientes que utilizam uma VPC ativem o tráfego de rede e os logs de DNS por meio dos Logs de fluxo de VPC e doslogs de consultas doAmazon Route 53 Resolver, respectivamente, transmitindo-os a um bucket do Amazon S3 ou a um grupo de logs do CloudWatch. É possível criar um log de fluxo de VPC, uma sub-rede ou uma interface de rede. Para logs de fluxo de VPC, é possível ser seletivo em relação a como e onde usar os logs de fluxo para reduzir o custo.

Logs do AWS CloudTrail, Logs de fluxo de VPC e logs de consulta do Route 53 Resolver são as fontes básicas de registro em log para oferecer compatibilidade com investigações de segurança na AWS. Também é possível usar o Amazon Security Lake para coletar, normalizar e armazenar esses dados de logs no formato do Apache Parquet e no Open Cybersecurity Schema Framework (OCSF), que estão prontos para consulta. O Security Lake também é compatível com outros logs da AWS e logs de fontes de terceiros.

Os serviços da AWS podem gerar logs não capturados pelas fontes de log básicas, como logs do Elastic Load Balancing, logs do AWS WAF, logs de gravador do AWS Config, descobertas do Amazon GuardDuty, logs de auditoria do Amazon Elastic Kubernetes Service (Amazon EKS) e logs de aplicações e do sistema de instâncias do Amazon EC2. Para ter uma lista completa de opções de registro em log e monitoramento, consulte Apêndice A: Definições de recursos de nuvem: registro em log e eventos do Guia de resposta a incidentes de segurança da AWS.

- Recursos de registro em log de pesquisa para cada serviço e aplicação da AWS: cada serviço e aplicação da AWS oferecem opções armazenamento de logs, sendo cada um com seus próprios recursos de retenção e ciclo de vida. Os dois serviços de armazenamento de logs mais comuns são Amazon Simple Storage Service (Amazon S3) e Amazon CloudWatch. Para períodos de retenção longos, é recomendável utilizar o Amazon S3 para seus recursos de economia e ciclo de vida flexíveis. Se a opção de registro em log principal for logs do Amazon CloudWatch, como opção, você deve considerar o arquivamento de logs menos acessados no Amazon S3.
- Selecione o armazenamento de logs: a escolha do armazenamento de logs, geralmente, é relacionada a qual ferramenta de consultas você utiliza, recursos de retenção, familiaridade e custo. As principais opções para armazenamento de logs são um bucket do Amazon S3 ou um grupo de logs do CloudWatch.

  Um bucket do Amazon S3 oferece armazenamento econômico e durável com uma política de ciclo de vida opcional. Os logs armazenados em buckets do Amazon S3 podem ser consultados com serviços como o Amazon Athena.

Um grupo de logs do CloudWatch oferece armazenamento durável e um recurso de consultas incorporado por meio do CloudWatch Logs Insights.

- Identifique a retenção de logs apropriada: quando você utiliza um bucket do Amazon S3 ou o grupo de logs do CloudWatch para armazenar logs, é necessário estabelecer ciclos de vida adequados para cada fonte de logs a fim de otimizar os custos de armazenamento e recuperação. Os clientes geralmente têm entre três meses a um ano de logs prontamente disponíveis para consultas, com retenção de até sete anos. A escolha de disponibilidade e retenção deve se alinhar aos seus requisitos de segurança e um composto de atribuições regulatórias, estatutárias e de negócios.
- Ative o registro em log para cada serviço e aplicação da AWS com políticas adequadas de retenção e ciclo de vida: para cada serviço ou aplicação da AWS em sua organização, procure as orientações específicas de configuração de registro em log:
  - Configurar a trilha do AWS CloudTrail
  - Configurar logs de fluxo de VPC
  - Configurar as exportações de descobertas do Amazon GuardDuty
  - Configurar os registros do AWS Config
  - Configurar o tráfego de ACL da web do AWS WAF
  - Configurar os logs de tráfego de rede do AWS Network Firewall
  - Configurar logs de acesso do Elastic Load Balancing
  - Configurar logs de consulta do Amazon Route 53 resolver
  - Configurar logs do Amazon RDS
  - Configurar logs do ambiente de gerenciamento Amazon EKS
  - Configurar o agente do Amazon CloudWatch para instâncias do Amazon EC2 e servidores on-premises
- Selecione e implemente os mecanismos de consulta para logs: para consultas de log, você pode usar o CloudWatch Logs Insights para dados armazenados em grupos de logs do CloudWatch, e o Amazon Athena e o Amazon OpenSearch Service para dados armazenados no Amazon S3. Também é possível usar ferramentas de consulta de terceiros, como um serviço de gerenciamento de eventos e informações de segurança (SIEM).

  O processo para selecionar uma ferramenta de consulta de log deve considerar as pessoas, o processo e os aspectos de tecnologia de suas operações de segurança. Selecione uma ferramenta que atenda aos requisitos operacionais, de negócios e segurança, esteja acessível

e possa receber manutenção no longo prazo. Lembre-se de que as ferramentas de consulta de logs funcionam da forma ideal quando o número de logs a serem verificados é mantido dentro dos limites da ferramenta. Não é incomum ter várias ferramentas de consulta devido a restrições financeiras ou técnicas.

Por exemplo, você pode usar uma ferramenta de gerenciamento de eventos e informações de segurança (SIEM) de terceiros para realizar consultas para os últimos 90 dias de dados, mas usar o Athena para realizar consultas além de 90 dias devido ao custo de ingestão de logs de um SIEM. Seja qual for a implementação, garanta que sua abordagem minimize o número de ferramentas necessárias para maximizar a eficiência operacional, especialmente durante a investigação de um evento de segurança.

- Use logs para alertas: a AWS oferece alertas por meio de vários serviços de segurança:
  - O 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.
  - O Amazon GuardDuty é um serviço de detecção de ameaças que monitora de forma contínua a existência de atividade mal-intencionada e comportamento não autorizado para proteger suas Contas da AWS e workloads. O GuardDuty ingere, agrega e analisa informações de fontes, como eventos de dados e gerenciamento do AWS CloudTrail, logs de DNS, logs de fluxo de VPC e logs do Amazon EKS Audit. O GuardDuty extrai fluxos de dados independentes diretamente do CloudTrail, de logs de fluxo de VPC, logs de consulta ao DNS e do Amazon EKS. Não é necessário gerenciar políticas de bucket do Amazon S3 nem modificar a forma de coletar e armazenar logs. Ainda é recomendável reter esses logs para sua própria investigação e fins de conformidade.
  - O 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.

  Você também pode utilizar mecanismos de geração de alertas personalizados para alertas de segurança não cobertos por esses serviços ou para alertas específicos relevantes para o seu ambiente. Para ter informações sobre a criação desses alertas e detecções, consulte Detecção no Guia de resposta a incidentes de segurança da AWS.

Recursos

Práticas recomendadas relacionadas:

- SEC04-BP02 Analisar logs, descobertas e métricas de forma centralizada

- SEC07-BP04 Definir o gerenciamento do ciclo de vida de dados
- SEC10-BP06 Pré-implantação de ferramentas

Documentos relacionados:

- Guia de resposta a incidentes de segurança da AWS
- Conceitos básicos do Amazon Security Lake
- Conceitos básicos: Amazon CloudWatch Logs
- Soluções de segurança parceiros: registro em log e monitoramento

Vídeos relacionados:

- AWS re:Invent 2022: Introdução ao Amazon Security Lake

Exemplos relacionados:

- Assisted Log Enabler for AWS
- Exportação histórica de descobertas do AWS Security Hub

Ferramentas relacionadas:

- Snowflake for Cybersecurity

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

- SEGURANÇA 5. Como proteger seus recursos de rede?
- SEGURANÇA 6. Como proteger seus recursos de computação?

SEGURANÇA 5. Como proteger 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 os componentes que compartilham requisitos de confidencialidade em camadas para minimizar o possível escopo do impacto do acesso não autorizado. 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/ou proveniente da Internet. O tráfego só deve fluir do próximo recurso menos sigiloso adjacente. Considere uma aplicação da web atrás de um balanceador de carga. Seu banco de dados não deve ser acessível diretamente do balanceador de carga. Somente a lógica de negócios ou o servidor da web tem acesso direto ao seu banco de dados.

Resultado desejado: criar uma rede em camadas. Redes em camadas ajudam a agrupar logicamente componentes de rede semelhantes. Elas também reduzem o possível escopo de impacto do acesso não autorizado à rede. Uma rede configurada adequadamente em camadas dificulta que usuários não autorizados adaptem recursos adicionais em seu ambiente da AWS. Além de garantir caminhos de rede internos, você também deve proteger sua borda de rede, como aplicações da web e endpoints de API.

Antipadrões comuns:

- Criar todos os recursos em uma única VPC ou sub-rede.
- Utilizar grupos de segurança excessivamente permissivos.
- Não utilizar sub-redes.
- Permitir o acesso direto aos armazenamentos de dados, como bancos de dados.

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

Orientação de implementação

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. Considere implantar workloads sem servidor, como funções do Lambda, em uma VPC ou atrás de um Amazon API Gateway. As tarefas do AWS Fargate (Fargate) que não têm necessidade de acesso à Internet devem ser colocadas em sub-redes sem rota para ou proveniente da Internet. Essa abordagem em camadas reduz o impacto da configuração incorreta de uma única camada, o que pode permitir o acesso não intencional. Para o AWS Lambda, você pode executar as funções em sua VPC para utilizar os controles baseados em VPC.

Para a conectividade de rede que pode incluir milhares de VPCs, Contas da AWS e redes on-premises, você deve utilizar o AWS Transit Gateway. O Transit Gateway age como um hub que controla como o tráfego é roteado entre todas as redes conectadas, que agem como raios. O tráfego entre o Amazon Virtual Private Cloud (Amazon VPC) e o Transit Gateway permanece na rede privada da AWS, o que reduz a exposição externa a usuários não autorizados e possíveis problemas de segurança. O emparelhamento entre regiões do Transit Gateway também criptografa o tráfego entre regiões sem nenhum ponto único de falha ou gargalo de largura de banda.

Etapas da implementação

- Utilize o Reachability Analyzer para analisar o caminho entre uma origem e um destino com base na configuração: o Reachability Analyzer permite a você automatizar a verificação da conectividade para e proveniente de recursos conectados à VPC. Observe que essa análise é realizada analisando a configuração (nenhum pacote de rede é enviado na realização da análise).
- Utilize o Analisador de Acesso à Rede Amazon VPC para identificar o acesso acidental à rede aos recursos: o Analisador de Acesso à Rede Amazon VPC possibilita especificar seus requisitos de acesso à rede identificar possíveis caminhos de rede.
- Considere se os recursos precisam estar em uma sub-rede pública: não coloque os recursos em sub-redes públicas de sua VPC a menos que eles devam receber tráfego de rede de entrada de origens públicas.
- Crie sub-redes em suas VPCs: crie sub-redes para cada camada de rede (em grupos que incluam várias zonas de disponibilidade) para melhorar a microssegmentação. Verifique também se você associou as tabelas de rotas corretas com suas sub-redes para controlar o roteamento e a conectividade de rede.
- Utilize o AWS Firewall Manager para gerenciar seus grupos de segurança de VPC: o AWS Firewall Manager ajuda a reduzir o trabalho de usar vários grupos de segurança.
- Utilize o AWS WAF para proteger contra vulnerabilidades comuns da web: o AWS WAF pode ajudar a melhorar a segurança de borda inspecionando o tráfego quanto a vulnerabilidades comuns da web, como injeção de SQL. Ele também permite restringir o tráfego de endereços IP originários de determinados países ou locais geográficos.

- Utilize o Amazon CloudFront como uma rede de distribuição de conteúdo (CDN): o Amazon CloudFront pode ajudar a acelerar sua aplicação da web armazenando dados mais perto de seus usuários. Ele também pode melhorar a segurança de borda aplicando HTTPS, restringindo o acesso a áreas geográficas e garantindo que o tráfego de rede possa acessar somente recursos roteados por meio do CloudFront.
- Utilize o Amazon API Gateway ao criar interfaces de programação de aplicações (APIs): o Amazon API Gateway ajuda a publicar, monitorar e proteger APIs REST, HTTPS e de WebSocket.

Recursos

Documentos relacionados:

- AWS Firewall Manager
- Amazon Inspector
- Segurança na Amazon VPC
- Reachability Analyzer
- Analisador de Acesso à Rede Amazon VPC

Vídeos relacionados:

- Arquiteturas de referência do AWS Transit Gateway para várias VPCs
- Aceleração e proteção de aplicações com o Amazon CloudFront, o AWS WAF e o AWS Shield
- AWS re:Inforce 2022: Validar controles de acesso à rede eficazes na AWS
- AWS re:Inforce 2022: Proteções avançadas contra bots usando o AWS WAF

Exemplos relacionados:

- Well-Architected Lab: Implantação automatizada de VPC
- Workshop: Analisador de Acesso à Rede Amazon 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

SEGURANÇA 6. Como proteger 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.

Resultado desejado: criar e manter um programa de gerenciamento de vulnerabilidade. Verificar regularmente e corrigir recursos, como instâncias do Amazon EC2, contêineres do Amazon Elastic

Container Service (Amazon ECS) e workloads do Amazon Elastic Kubernetes Service (Amazon EKS). Configurar janelas de manutenção para recursos gerenciados da AWS, como bancos de dados Amazon Relational Database Service (Amazon RDS). Utilizar a verificação de código estático para inspecionar a existência de problemas comuns no código-fonte da aplicação. Considerar testes de penetração de aplicações da web se sua organização tiver as habilidades obrigatórias ou puder contratar assistência externa.

Antipadrões comuns:

- Não ter um programa de gerenciamento de vulnerabilidades.
- Realizar aplicação de patches do sistema sem considerar a gravidade ou como evitar riscos.
- Utilizar software que ultrapassou a data de fim de vida útil (EOL) indicada pelo fornecedor.
- Implantar código em produção antes de analisar a existência de problemas de segurança.

Benefícios do estabelecimento desta prática recomendada:

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

Orientação de implementação

Um programa de gerenciamento de vulnerabilidades inclui avaliação de segurança, identificação de problemas, priorização e realização de operações de patch como parte da solução dos problemas. A automação é a chave para verificar de forma contínua as workloads quanto a problemas e exposição acidental à rede e realização de remediação. Automatizar a criação e atualizar os recursos economiza tempo e reduz o risco de erros de configuração que criam mais problemas. Um programa de gerenciamento de vulnerabilidades bem projetado também deve considerar testes de vulnerabilidades durante o desenvolvimento e os estágios de implantação do ciclo de vida do software. Implementar o gerenciamento de vulnerabilidades durante o desenvolvimento e a implantação ajuda a reduzir a chance de uma vulnerabilidade atingir seu ambiente de produção.

Implementar um programa de gerenciamento de vulnerabilidades exige um bom entendimento do Modelo de responsabilidade compartilhada da AWS e como ele se relaciona com suas workloads específicas. Segundo o modelo de responsabilidade compartilhada, a AWS é responsável por proteger a infraestrutura da Nuvem AWS. Essa infraestrutura abrange o hardware, o software, as redes e as instalações que executam os serviços da Nuvem AWS. Você é responsável pela segurança na nuvem, por exemplo, os dados reais, a configuração de segurança e as tarefas de gerenciamento de instâncias do Amazon EC2 e por garantir que seus objetos do Amazon S3 sejam

classificados e configurados corretamente. Sua abordagem ao gerenciamento de vulnerabilidades também pode variar dependendo dos serviços consumidos. Por exemplo, a AWS gerencia a aplicação de patches para nosso serviço de banco de dados relacional gerenciado, o Amazon RDS, mas você seria responsável pela colocação de patches em bancos de dados auto-hospedados.

A AWS tem uma série de serviços para ajudar com seu programa de gerenciamento de vulnerabilidades. O Amazon Inspector verifica de forma contínua as workloads da AWS quanto a problemas de software e acesso acidental à rede. O AWS Systems Manager Patch Manager ajuda a gerenciar a aplicação de patches em suas instâncias do Amazon EC2. O Amazon Inspector e o Systems Manager podem ser visualizados no AWS Security Hub, um serviço de gerenciamento de procedimentos de segurança na nuvem que ajuda a automatizar verificações de segurança da AWS e centralizar alertas de segurança.

O Amazon CodeGuru pode ajudar a identificar possíveis problemas em aplicações Java e Python utilizando análise de código estático.

Etapas da implementação

- Configurar o Amazon Inspector: o Amazon Inspector detecta automaticamente instâncias do Amazon EC2 recém-executadas, funções do Lambda e imagens de contêiner elegíveis enviadas ao Amazon ECR e as verifica imediatamente quanto a problemas de software, possíveis defeitos e exposição acidental à rede.
- Verificar o código-fonte: verifique as bibliotecas e as dependências quanto a problemas e defeitos. O Amazon CodeGuru pode verificar e fornecer recomendações para corrigir problemas de segurança comuns para aplicações Java e Python. A OWASP Foundation publica uma lista de ferramentas de análise de código-fonte (também conhecidas como ferramentas SAST).
- Implementar um mecanismo para verificar e aplicar patches ao seu ambiente existente, bem como verificação como parte de um processo de construção de pipeline de CI/CD: implemente um mecanismo para verificar e aplicar patches quanto a problemas em suas dependências e sistemas operacionais a fim de ajudar a proteger-se contra novas ameaças. Execute esse mecanismo regularmente. O gerenciamento de vulnerabilidade de software é essencial ao entendimento de onde é necessário aplicar patches ou resolver problemas de software. Priorize a remediação de possíveis problemas de segurança incorporando avaliações de vulnerabilidade no início de seu pipeline de integração/entrega contínua (CI/CD). Sua abordagem pode variar com base nos serviços da AWS que você está consumindo. Para conferir a existência de possíveis problemas no software em execução em instâncias do Amazon EC2, adicione o Amazon Inspector ao seu pipeline para alertar e interromper o processo de compilação se forem detectados problemas ou possíveis defeitos. O Amazon Inspector monitora recursos de forma contínua. Você também

pode utilizar produtos de código aberto, como OWASP Dependency-Check, Snyk, OpenVAS, gerenciadores de pacotes e ferramentas de AWS Partner para gerenciamento de vulnerabilidades.

- Utilize o AWS Systems Manager: 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 outros recursos de computação. O AWS Systems Manager Patch Manager automatiza o processo de aplicação de patches em instâncias gerenciadas com atualizações relacionadas à segurança e outros tipos de atualizações. O Patch Manager pode ser utilizado para aplicar patches em instâncias do Amazon EC2 para sistemas operacionais e aplicações, como aplicações da Microsoft, pacotes de serviços Windows e atualizações de versão secundária para instâncias baseadas em Linux. Além do Amazon EC2, o Patch Manager também pode ser utilizado para aplicar patches em servidores on-premises.

  Para ter uma lista de sistemas operacionais compatíveis, consulte Sistemas operacionais compatíveis no Guia do usuário do Systems Manager. Você pode verificar instâncias para ver apenas um relatório de patches ausentes ou verificar e instalar automaticamente todos os patches ausentes.
- Utilize do AWS Security Hub: o Security Hub oferece uma visão abrangente do estado de seu sistema na AWS. Ele coleta dados de segurança em vários serviços da AWS e oferece essas descobertas em um formato personalizado, possibilitando priorizar as descobertas de segurança em serviços da AWS.
- Utilize o AWS CloudFormation: o AWS CloudFormation é um serviço de infraestrutura como código (IaC) que pode ajudar com o gerenciamento de vulnerabilidades automatizando a implantação de recursos e padronizando a arquitetura de recursos em várias contas e ambientes.

Recursos

Documentos relacionados:

- AWS Systems Manager
- Visão geral de segurança do AWS Lambda
- Amazon CodeGuru
- Gerenciamento aprimorado e automatizado de vulnerabilidades para workloads de nuvem com um novo Amazon Inspector
- Automatizar o gerenciamento e a remediação de vulnerabilidades na AWS usando o Amazon Inspector e o AWS Systems Manager: Parte 1

Vídeos relacionados:

- Proteção de serviços com tecnologia sem servidor e de contêiner
- Práticas recomendadas de segurança para o serviço de metadados de instância do Amazon EC2

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

- SEGURANÇA 7. Como classificar meus dados?
- SEGURANÇA 8. Como proteger seus dados em repouso?
- SEGURANÇA 9. Como proteger seus dados em trânsito?

SEGURANÇA 7. Como classificar meus dados?

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

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

É essencial compreender o tipo e a classificação de dados que sua workload está processando, os processos de negócios associados, onde os dados são armazenados e quem é o proprietário dos dados. Você também deve ter uma compreensão dos requisitos legais e de conformidade aplicáveis de sua workload e quais controles de dados precisam ser implementados. A identificação dos dados é a primeira etapa da jornada da classificação de dados.

Benefícios do estabelecimento desta prática recomendada:

A classificação dos dados possibilita que os proprietários da workload identifiquem os locais que armazenam dados sigilosos e determinem como esses dados devem ser acessados e compartilhados.

A classificação dos dados tem como objetivo responder às seguintes perguntas:

- Qual tipo de dados você tem?

  Podem ser dados como:

  - Propriedade intelectual (IP), como segredos comerciais, patentes ou contratos.
  - Informações de saúde protegidas (PHI), como registros médicos que contêm informações do histórico médico referente a um indivíduo.
  - Informações de identificação pessoal (PII), como nome, endereço, data de nascimento e ID nacional ou número de registro.
  - Dados do cartão de crédito, como o Número da conta principal (PAN), nome do titular do cartão, data de validade e número do código de serviço.
  - Onde os dados sigilosos são armazenados?
  - Quem pode acessar, modificar e excluir dados?
  - A compreensão das permissões do usuário é essencial na proteção contra o possível uso indevido de dados.
- Quem pode realizar operações de criação, leitura, atualização e exclusão (CRUD)?
  - Considere a possível escalação de privilégios compreendendo quem pode gerenciar permissões aos dados.
- Qual impacto nos negócios poderá ocorrer se os dados forem divulgados de forma acidental, alterados ou excluídos?
  - Entenda a consequência do risco se os dados forem modificados, excluídos ou divulgados acidentalmente.

Ao responder a estas perguntas, você pode realizar as seguintes ações:

- Reduzir o escopo de dados sigilosos (como o número de locais de dados sigilosos) e limitar o acesso aos dados sigilosos somente para usuários aprovados.

- Obtenha um entendimento de diferentes tipos de dados para que você possa implementar técnicas e mecanismos de proteção de dados apropriados, como criptografia, prevenção da perda de dados e gerenciamento de identidade e acesso.
- Otimize os custos entregando os objetivos de controle certos para os dados.
- Responda às perguntas de modo confidencial de reguladores e auditores sobre os tipos e a quantidade de dados e como os dados de diferentes níveis de confidencialidade são isolados uns dos outros.

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

Orientação de implementação

Classificação dos dados é o ato de identificar a confidencialidade dos dados. Ela pode envolver marcação para tornar os dados facilmente acessíveis e rastreáveis. A classificação de dados também reduz a duplicação de dados, o que pode ajudar a reduzir os custos de armazenamento e backup enquanto acelera o processo de pesquisa.

Utilize serviços, como o Amazon Macie para automatizar em grande escala a descoberta e a classificação de dados sigilosos. Outros serviços, como Amazon EventBridge e AWS Config, podem ser utilizados para automatizar a remediação de problemas de segurança de dados, como buckets do Amazon Simple Storage Service (Amazon S3) não criptografados e volumes do Amazon EC2 EBS ou recursos de dados não marcados. Para ter uma lista completa de integrações de serviços da AWS, consulte a documentação do EventBridge.

A detecção de PII em dados não estruturados, como e-mails de clientes, tickets de suporte, análises de produtos e redes sociais, é possível com o uso do Amazon Comprehend, que é um serviço de processamento de linguagem natural (PLN) que utiliza machine learning (ML) para encontrar insights e relacionamentos, como pessoas, locais, sentimentos e tópicos em texto não estruturado. Para ter uma lista de serviços da AWS que podem auxiliar na identificação dos dados, consulte Técnicas comuns para detectar dados PHI e PII com o uso de serviços da AWS.

Outro método compatível com a classificação e a proteção de dados é a marcação de recursos da AWS. A marcação possibilita atribuir metadados aos seus recursos da AWS que você pode utilizar para gerenciar, identificar, organizar, procurar e filtrar recursos.

Em alguns casos, você pode optar por marcar recursos inteiros (como um bucket do S3), especialmente quando uma workload ou um serviço específico deve armazenar processos ou transmissões de classificação de dados já conhecidos.

Quando apropriado, é possível marcar um bucket do S3 em vez de objetos individuais para facilidade de administração e manutenção de segurança.

Etapas da implementação

Detectar dados sigilosos no Amazon S3:

1. Antes de começar, você deve ter permissões apropriadas para acessar o console do Amazon Macie e as operações de API. Para ter detalhes adicionais, consulte Conceitos básicos do Amazon Macie.
2. Utilize o Amazon Macie para realizar a descoberta de dados automatizada quando seus dados sigilosos residem no Amazon S3.
   - Utilize o guia Conceitos básicos do Amazon Macie para configurar um repositório para os resultados da descoberta de dados sigilosos e criar um trabalho de descoberta de dados sigilosos.
   - Como utilizar o Amazon Macie para visualizar dados sigilosos em buckets do S3.

     Por padrão, o Macie analisa objetos utilizando o conjunto de identificadores de dados gerenciados que recomendamos para a descoberta automatizada de dados sigilosos. É possível ajustar a análise configurando o Macie para utilizar identificadores de dados gerenciados específicos, identificadores de dados personalizados e listas de permissões quando ele realiza a descoberta automatizada de dados sigilosos para a sua conta ou organização. Você pode ajustar o escopo da análise excluindo buckets específicos (por exemplo, buckets do S3 que geralmente armazenam dados de registro em log da AWS).
3. Para configurar e utilizar a descoberta automatizada de dados sigilosos, consulte Realizar a descoberta automatizada de dados sigilosos com o Amazon Macie.
4. Você também pode considerar Descoberta automatizada de dados para o Amazon Macie.

Detectar dados sigilosos no Amazon RDS:

Para ter mais informações sobre a descoberta de dados em bancos de dados Amazon Relational Database Service (Amazon RDS), consulte Habilitar a classificação de dados para o banco de dados Amazon RDS com o Macie.

Detectar dados sigilosos no DynamoDB:

- Detectar dados sigilosos no DynamoDB com Macie explica como utilizar o Amazon Macie para detectar dados sigilosos em tabelas do Amazon DynamoDB exportando os dados para o Amazon S3 para verificação.

Soluções de parceiros da AWS:

- Considere utilizar nossa AWS Partner Network extensiva. Os parceiros da AWS têm ferramentas e frameworks de conformidade extensas que se integram diretamente aos serviços da AWS. Os parceiros podem oferecer uma solução de governança e conformidade personalizada para ajudar você a atender às suas necessidades organizacionais.
- Para saber sobre as soluções personalizadas em classificação de dados, consulte Governança de dados na era dos requisitos de regulamento e conformidade.

É possível aplicar automaticamente os padrões de marcação que sua organização adota criando e implantando políticas com o uso do AWS Organizations. As políticas de tag possibilitam especificar regras que definem nomes de chave válidas e quais valores são válidos para cada chave. É possível optar somente por monitorar, que oferece a você uma oportunidade de avaliar e limpar suas tags existentes. Depois que suas tags estiverem em conformidade com seus padrões escolhidos, você poderá ativar a aplicação nas políticas de tag a fim de impedir que tags sem conformidade sejam criadas. Para ter mais detalhes, consulte Proteger tags de recursos utilizadas para autorização utilizando uma política de controle de serviço no AWS Organizations e o exemplo de política em Impedir que as tags sejam modificadas, exceto por principais autorizados.

- Para começar a utilizar políticas de tag no AWS Organizations, é altamente recomendável que você siga o fluxo de trabalho em Conceitos básicos de políticas de tag antes de passar para políticas de tag mais avançadas. Compreender os efeitos de anexar uma política de tag simples a uma conta antes de expandir para uma unidade organizacional (UO) ou uma organização inteira permite ver os efeitos de uma política de tag antes de aplicar a conformidade com a política de tag. Conceitos básicos de políticas de tag oferece links para instruções de tarefas relacionadas a política mais avançadas.
- Considere a avaliação de outros serviços e recursos do AWS compatíveis com a classificação de dados, que estão listados no whitepaper Classificação de dados.

Recursos

Documentos relacionados:

- Conceitos básicos do Amazon Macie
- Descoberta automatizada de dados com o Amazon Macie
- Conceitos básicos de políticas de tag
- Detectar entidades de PII

Blogs relacionados:

- Como utilizar o Amazon Macie para visualizar dados sigilosos em buckets do S3.
- Realizar a descoberta automatizada de dados sigilosos com o Amazon Macie
- Técnicas comuns para detectar dados PHI e PII com o uso de serviços da AWS
- Detectar e editar PII com o uso do Amazon Comprehend
- Proteger tags de recursos usadas para autorização utilizando uma política de controle de serviços no AWS Organizations
- Habilitar a classificação do banco de dados Amazon RDS com o Macie
- Detectar dados sigilosos no DynamoDB com o Macie
-

Vídeos relacionados:

- Segurança dos dados orientada a eventos com o uso do Amazon Macie
- Amazon Macie para proteção e governança de dados
- Ajustar descobertas de dados sigilosos com listas de permissão

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)

SEGURANÇA 8. Como proteger 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

O gerenciamento seguro de chaves inclui o armazenamento, a rotação, o controle de acesso e o monitoramento do material essencial necessário para proteger os dados em repouso para sua workload.

Resultado desejado: um mecanismo de gerenciamento de chaves escalável, repetível e automatizado. O mecanismo deve fornecer a capacidade de impor o acesso de privilégio mínimo ao material essencial e fornecer o equilíbrio correto entre disponibilidade, confidencialidade e integridade das chaves. O acesso às chaves deve ser monitorado e o material da chave deve ser rotacionado por meio de um processo automatizado. O material de chave nunca deve estar acessível a identidades humanas.

Antipadrões comuns:

- Acesso humano a material de chave não criptografado.
- Criação de algoritmos criptográficos personalizados.
- Permissões excessivamente amplas para acessar materiais importantes.

Benefícios de estabelecer esta prática recomendada: Ao estabelecer um mecanismo seguro de gerenciamento de chaves para sua workload, você pode ajudar a proteger seu conteúdo contra acesso não autorizado. Além disso, você pode estar sujeito aos requisitos regulamentares para criptografar seus dados. Uma solução eficaz de gerenciamento de chaves pode fornecer mecanismos técnicos alinhados a essas regulamentações para proteger o material chave.

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

Orientação para implementação

Muitos requisitos regulatórios e práticas recomendadas incluem a criptografia de dados em repouso como um controle de segurança fundamental. Para cumprir esse controle, sua workload precisa de um mecanismo para armazenar e gerenciar com segurança o material chave usado para criptografar seus dados em repouso.

A AWS oferece o AWS Key Management Service (AWS KMS) para fornecer armazenamento durável, seguro e redundante para chaves do AWS KMS. Muitos serviços da AWS se integram com o AWS KMS para oferecer suporte à criptografia de seus dados. O AWS KMS usa módulos de segurança de hardware validados pelo FIPS 140-2 Nível 3 para proteger suas chaves. Não há mecanismo para exportar chaves do AWS KMS em texto simples.

Ao implantar workloads usando uma estratégia de várias contas, isso é considerado prática recomendada para manter chaves do AWS KMS na mesma conta da workload que as usa. Nesse modelo distribuído, a responsabilidade pelo gerenciamento das chaves do AWS KMS é da equipe de aplicações. Em outros casos de uso, as organizações podem optar por armazenar as chaves do AWS KMS em uma conta centralizada. Essa estrutura centralizada requer políticas adicionais para permitir o acesso entre contas necessário para que a conta da workload acesse as chaves armazenadas na conta centralizada, mas pode ser mais aplicável em casos de uso em que uma única chave é compartilhada entre várias Contas da AWS.

Independentemente de onde o material da chave esteja armazenado, o acesso à chave deve ser rigorosamente controlado por meio do uso de políticas de chave e políticas do IAM. Políticas de chave são a principal forma de controlar o acesso a uma chave do AWS KMS. Além disso, concessões à chave do AWS KMS podem fornecer acesso a serviços da AWS para criptografar e descriptografar dados em seu nome. Reserve um tempo para revisar as práticas recomendadas para controle de acesso às chaves do AWS KMS.

É uma prática recomendada monitorar o uso de chaves de criptografia para detectar padrões de acesso incomuns. As operações realizadas usando chaves gerenciadas pela AWS e chaves gerenciadas pelo cliente armazenadas no AWS KMS podem ser registradas no AWS CloudTrail e devem ser revisadas periodicamente. Atenção especial deve ser dada ao monitoramento dos principais eventos de destruição. Para mitigar a destruição acidental ou maliciosa de material de chave, os eventos de destruição da chave não excluem o material da chave imediatamente. As tentativas de excluir as chaves no AWS KMS estão sujeitas a um período de espera, cujo padrão é

de 30 dias, dando aos administradores tempo para revisar essas ações e reverter a solicitação, se necessário.

A maioria dos serviços da AWS usam o AWS KMS de forma transparente para você. Seu único requisito é decidir se quer usar uma chave gerenciada pela AWS ou gerenciada pelo cliente. Se sua workload exigir o uso direto de AWS KMS para criptografar ou descriptografar dados, a prática recomendada é usar criptografia envelopada para proteger seus dados. O SDK de criptografia da AWS pode fornecer às suas aplicações primitivas de criptografia do lado do cliente para implementar a criptografia envelopada e integrar com o AWS KMS.

Etapas da implementação

1. Determine as opções adequadas de gerenciamento de chaves (gerenciado pela AWS ou gerenciado pelo cliente) para a chave.
   - Para facilitar o uso, a AWS oferece, para a maioria dos serviços, chaves de propriedade da AWS e gerenciadas por ela, que fornecem capacidade de criptografia em repouso sem a necessidade de gerenciar materiais ou políticas de chaves.
   - Ao usar chaves gerenciadas pelo cliente, considere o armazenamento de chaves padrão para fornecer o melhor equilíbrio entre agilidade, segurança, soberania de dados e disponibilidade. Outros casos de uso podem exigir o uso de armazenamentos de chaves personalizadas com AWS CloudHSM ou o armazenamento de chaves externo.
2. Analise a lista de serviços que você está usando para sua workload para entender como o AWS KMS se integra ao serviço. Por exemplo, as instâncias do EC2 podem usar volumes criptografados do EBS, verificando se os snapshots do Amazon EBS criados a partir desses volumes também são criptografados usando uma chave gerenciada pelo cliente e mitigando a divulgação acidental de dados de snapshots não criptografados.
   - Como os serviços da AWS são usados no AWS KMS
   - Para obter informações detalhadas sobre as opções de criptografia que um serviço da AWS oferece, consulte o tópico Criptografia em repouso no guia do usuário ou no guia do desenvolvedor do serviço.
3. Implemente o AWS KMS: o AWS KMS simplifica a criação e o gerenciamento de chaves e o controle do uso da criptografia em uma ampla variedade de serviços da AWS e em suas aplicações.
   - Conceitos básicos: AWS Key Management Service (AWS KMS)
   - Revise as práticas recomendadas para controle de acesso às chaves do AWS KMS.

4. Considere o AWS Encryption SDK: use a integração do AWS Encryption SDK com o AWS KMS quando sua aplicação precisar criptografar dados no lado do cliente.
   - AWS Encryption SDK
5. Habilite o IAM Access Analyzer para revisar e notificar automaticamente se houver políticas de chave do AWS KMS excessivamente amplas.
6. Habilite o Security Hub para receber notificações se houver políticas de chaves configuradas incorretamente, chaves programadas para exclusão ou chaves sem a rotação automática ativada.
7. Determine o nível de registro em log apropriado para suas chaves do AWS KMS. Como as chamadas para o AWS KMS, incluindo eventos somente para leitura, são registradas em log, os logs do CloudTrail associados ao AWS KMS podem se tornar volumosos.
   - Algumas organizações preferem registrar a atividade de log do AWS KMS em uma trilha separada. Para obter mais detalhes, consulte a seção Registro em log de chamadas de API do AWS KMS com CloudTrail do guia do desenvolvedor do AWS KMS.

Recursos

Documentos relacionados:

- AWS Key Management Service
- Ferramentas e serviços criptográficos da AWS
- Como proteger dados do Amazon S3 com o uso de criptografia
- Criptografia envelopada
- Promessa de soberania digital
- Desmistificação das operações de chave do AWS KMS, traga sua própria chave, armazenamento de chaves personalizado e portabilidade de texto cifrado
- Detalhes criptográficos do AWS Key Management Service

Vídeos relacionados:

- How Encryption Works in AWS (Como funciona a criptografia na AWS)
- Securing Your Block Storage on AWS (Proteger seu armazenamento em bloco na AWS)
- AWS data protection: Using locks, keys, signatures, and certificates (Proteção de dados da AWS: uso de travas, chaves, assinaturas e certificados)

Exemplos relacionados:

- Implemente mecanismos avançados de controle de acesso usando o AWS KMS

SEC08-BP02 Aplicar criptografia em repouso

É necessário implementar o uso de criptografia para dados em repouso. A criptografia mantém a confidencialidade dos dados sigilosos em caso de acesso não autorizado ou divulgação acidental.

Resultado desejado: os dados privados devem ser criptografados por padrão quando em repouso. A criptografia ajuda a manter a confidencialidade dos dados e oferece uma camada adicional de proteção contra a divulgação intencional ou acidental de dados ou exfiltração. Dados criptografados não podem ser lidos nem acessados sem primeiro descriptografá-los. Todos os dados armazenados não criptografados devem ser inventariados e controlados.

Antipadrões comuns:

- Não utilizar configurações de criptografia por padrão.
- Conceder acesso excessivamente permissivo para chaves de descriptografia.
- Não monitorar o uso de chaves de criptografia e descriptografia.
- Armazenar dados não criptografados.
- Utilizar a mesma chave de criptografia para todos os dados, seja qual for o uso, os tipos e a classificação de dados.

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

Orientação de implementação

Mapeie as chaves de criptografia às classificações de dados em suas workloads. Essa abordagem ajuda a proteger-se contra o acesso excessivamente permissivo ao utilizar uma chave única ou um número muito pequeno de chaves de criptografia para seus dados (consulte SEC07-BP01 Identificar os dados em sua workload).

O AWS Key Management Service (AWS KMS) integra-se a muitos serviços da AWS para facilitar a criptografia de seus dados em repouso. Por exemplo, no Amazon Simple Storage Service (Amazon S3), você pode definir a criptografia padrão em um bucket para que os novos objetos sejam criptografados automaticamente. Ao utilizar o AWS KMS, considere o nível de restrição necessário

para os dados. Chaves do AWS KMS controladas por serviço e padrão são gerenciadas e utilizadas em seu nome pelo AWS. Para dados sigilosos que exijam acesso refinado à chave de criptografia subjacente, considere chaves gerenciadas pelo cliente (CMKs). Você tem total controle sobre as CMKs, como gerenciamento de alternância e acesso pelo uso de políticas de chave.

Além disso, o Amazon Elastic Compute Cloud (Amazon EC2) e o Amazon S3 são compatíveis com a imposição de criptografia ao definir a criptografia padrão. Você pode usar o Regras do AWS Config para conferir automaticamente se está usando criptografia, por exemplo, volumes do Amazon Elastic Block Store (Amazon EBS), instâncias do Amazon Relational Database Service (Amazon RDS) e buckets do Amazon S3.

A AWS também oferece operações de criptografia do lado do cliente, possibilitando que você criptografe os dados antes de fazer upload deles para a nuvem. O AWS Encryption SDKoferece uma forma de criptografar seus dados utilizando a criptografia envelopada. Você fornece a chave de encerramento, e o AWS Encryption SDK gera uma chave de dados exclusiva para cada objeto de dados que ele criptografa. Considere utilizar o AWS CloudHSM se precisar de um módulo de segurança de hardware de um locatário (HSM) gerenciado. O AWS CloudHSM possibilita gerar, importar e gerenciar chaves criptográficas em um HSM validado de nível 3 FIPS 140-2. Alguns casos de uso do AWS CloudHSM incluem proteger chaves privadas para emitir uma autoridade de certificado (CA) e ativar a criptografia de dados transparente (TDE) para bancos de dados Oracle. O AWS CloudHSM Client SDK oferece software que possibilita criptografar dados do lado do cliente com chaves armazenadas no AWS CloudHSM antes de fazer upload de seus dados para AWS. O Amazon DynamoDB Encryption Client também possibilita criptografar e assinar itens antes de fazer upload para uma tabela do DynamoDB.

Etapas da implementação

- Impor criptografia em repouso para o Amazon S3: implemente a criptografia padrão do bucket do Amazon S3.

  Configurar a criptografia padrão para volumes do Amazon EBS: especifique que você deseja que todos os volumes do Amazon 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.

  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.

  Configurar a criptografia do Amazon RDS: configure a criptografia para seus clusters de banco de dados Amazon RDS e snapshots em repouso utilizando a opção de criptografia.

Criar e configurar chaves do AWS KMS com políticas que limitem o acesso às entidades principais apropriadas para cada classificação de dados: por exemplo, crie uma chave do AWS KMS para criptografar dados de produção e uma chave diferente para criptografar dados de desenvolvimento ou teste. Você também pode conceder acesso de chave a outras Contas da AWS. Considere ter contas diferentes para seus ambientes de desenvolvimento e produção. Se seu ambiente de produção precisar descriptografar artefatos na conta de desenvolvimento, você poderá editar a política de CMK utilizada para criptografar os artefatos de desenvolvimento a fim de conferir à conta de produção a capacidade de descriptografar esses artefatos. O ambiente de produção pode, então, ingerir os dados descriptografados para uso na produção.

Configurar a criptografia em serviços da AWS adicionais: para outros serviços da AWS utilizados, leia a documentação de segurança do serviço em questão para determinar as opções de criptografia do serviço.

Recursos

Documentos relacionados:

- Ferramentas de criptografia da AWS
- Documentação da AWS
- AWS Encryption SDK
- Whitepaper de detalhes criptográficos do AWS KMS
- AWS Key Management Service
- Ferramentas e serviços criptográficos da AWS
- Criptografia do Amazon EBS
- Criptografia padrão de volumes do Amazon EBS
- Criptografar recursos do Amazon RDS
- Como ativo a criptografia padrão para um bucket do Amazon S3?
- Proteger dados do Amazon S3 com o uso de criptografia

Vídeos relacionados:

- Como a criptografia funciona na AWS
- 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

Para ajudar a proteger seus dados em repouso, implemente o controle de acesso utilizando mecanismos, como isolamento e versionamento, e aplique o princípio de privilégio mínimo. Evite conceder acesso público aos seus dados.

Resultado desejado: garantir que somente usuários autorizados possam acessar os dados conforme a necessidade. Proteger seus dados com backups regulares e versionamento a fim de impedir a modificação ou a exclusão de dados intencionais ou acidentais. Isolar dados críticos de outros dados a fim de proteger a confidencialidade e a integridade deles.

Antipadrões comuns:

- Armazenar dados com requisitos de confidencialidade ou classificações diferentes juntos.
- Utilizar permissões excessivamente permissivas em chaves de descriptografia.
- Classificar dados de modo inadequado.
- Não reter backups detalhados de dados importantes.
- Conceder acesso persistente a dados de produção.
- Não auditar o acesso aos dados nem rever as permissões regularmente

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

Orientação de implementação

Vários controles podem ajudar a proteger seus dados em repouso, por exemplo, acesso (utilizando privilégio mínimo), isolamento e versionamento. Deve ser feita a auditoria de acesso aos seus dados com os mecanismos de detecção, como AWS CloudTrail e os logs 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 criar um plano para reduzir a quantidade de dados disponíveis ao longo do tempo.

O Amazon S3 Glacier Vault Lock e o Amazon S3 Object Lock fornecem controle de acesso obrigatório para os objetos no Amazon S3. 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.

Etapas da implementação

- Aplicar o controle de acesso: aplique o controle de acesso com privilégios mínimos, incluindo acesso a chaves de criptografia.

- Dados separados com base em diferentes níveis de classificação: use diferentes Contas da AWS para níveis de classificação de dados e gerencie essas contas com o AWS Organizations.
- Analisar as políticas do AWS Key Management Service (AWS KMS): analise o nível de acesso concedido nas políticas do AWS KMS.
- Revisar as permissões de objeto e de bucket do Amazon S3: revise regularmente o nível de acesso concedido nas políticas de bucket do S3. Uma das práticas recomendadas é evitar 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. Garanta que os buckets que não devem permitir acesso público sejam configurados adequadamente para evitar o acesso público. Por padrão, todos os buckets do S3 são privados e só ser acessados por usuários que receberam explicitamente esse acesso.
- Ativar o AWS IAM Access Analyzer: o IAM Access Analyzer analisa os buckets do Amazon S3 e gera uma descoberta quando uma política do S3 concede acesso a uma entidade externa.
- Habilitar o versionamento do Amazon S3 e o bloqueio de objetos quando apropriado.
- Utilizar o Amazon S3 Inventory: o Amazon S3 Inventory pode ser usado para auditar e gerar relatórios sobre o status de replicação e criptografia de seus objetos do S3.
- Revisar as permissões do Amazon EBS e do compartilhamento de AMIs: as permissões de compartilhamento podem permitir que imagens e volumes sejam compartilhados com Contas da AWS externas à sua workload.
- Revise os compartilhamentos do AWS Resource Access Manager periodicamente para determinar se os recursos devem continuar a ser compartilhados. O Resource Access Manager possibilita compartilhar recursos, como políticas do AWS Network Firewall, regras do Amazon Route 53 Resolver e sub-redes em suas Amazon VPCs. Faça auditoria em recursos compartilhados regularmente e interrompa o compartilhamento dos que não precisem mais ser compartilhados.

Recursos

Práticas recomendadas relacionadas:

- SEC03-BP01 Definir requisitos de acesso
- SEC03-BP02 Conceder acesso com privilégio mínimo

Documentos relacionados:

- Whitepaper de detalhes criptográficos do AWS KMS

- Introdução ao gerenciamento de permissões de acesso aos seus recursos do Amazon S3
- Visão geral do gerenciamento de acesso dos recursos do AWS KMS
- Regras do AWS Config
- Amazon S3 + Amazon CloudFront: uma combinação perfeita na nuvem
- Usar versionamento
- Bloquear objetos usando o Bloqueio de objetos do Amazon S3
- Compartilhar um snapshot do Amazon EBS
- AMIs compartilhadas
- Hospedar uma aplicação de uma página no Amazon S3

Vídeos relacionados:

- 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)

SEGURANÇA 9. Como proteger 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 Aplicar 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

Os certificados Transport Layer Security (TLS) são usados para proteger as comunicações de rede e estabelecer a identidade de sites, recursos e workloads na internet, bem como em redes privadas.

Resultado desejado: Um sistema seguro de gerenciamento de certificados que pode provisionar, implantar, armazenar e renovar certificados em uma infraestrutura de chave pública (PKI). Um mecanismo seguro de gerenciamento de chaves e certificados evita que o material da chave privada do certificado seja divulgado e renova automaticamente o certificado periodicamente. Ele também se integra a outros serviços para fornecer comunicações de rede seguras e identidade para os recursos da máquina na workload. O material de chave nunca deve estar acessível a identidades humanas.

Antipadrões comuns:

- Executar etapas manuais durante os processos de implantação ou renovação de certificado.
- Não prestar a devida atenção à hierarquia da autoridade de certificação (CA) ao criar uma CA privada.
- Usar certificados autoassinados para recursos públicos.

Benefícios de estabelecer esta prática recomendada:

- Simplificar o gerenciamento de certificados por meio de implantação e renovação automatizadas.
- Incentivar a criptografia de dados em trânsito usando certificados TLS.
- Aumentar a segurança e a auditabilidade das ações de certificação realizadas pela autoridade de certificação.
- Organizar as tarefas de gerenciamento em diferentes camadas da hierarquia da CA.

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

Orientação para implementação

As workloads modernas fazem uso extensivo de comunicações de rede criptografadas usando protocolos de PKI, como TLS. O gerenciamento de certificados PKI pode ser complexo, mas o provisionamento, a implantação e a renovação automatizados de certificados podem reduzir o atrito associado ao gerenciamento deles.

A AWS oferece dois serviços para gerenciar certificados de PKI de uso geral: AWS Certificate Manager e AWS Private Certificate Authority (AWS Private CA). O ACM é o principal serviço que os clientes usam para provisionar, gerenciar e implantar certificados para uso em workloads públicas e privadas da AWS. O ACM emite certificados usando o AWS Private CA e integra-se a muitos outros serviços gerenciados da AWS para fornecer certificados TLS seguros para workloads.

A AWS Private CA permite estabelecer a própria autoridade de certificação raiz ou subordinada e emitir certificados TLS por meio de uma API. É possível usar esses tipos de certificado em cenários em que você controla e gerencia a cadeia de confiança do lado do cliente da conexão TLS. Além dos casos de uso do TLS, a AWS Private CA pode ser usada para emitir certificados para pods do Kubernetes, atestados de produtos de dispositivos Matter, assinatura de código e outros casos de uso com um modelo personalizado. Você também pode usar IAM Roles Anywhere para fornecer credenciais do IAM temporárias para workloads on-premises que receberam certificados X.509 assinados pela CA privada.

Além do ACM e do AWS Private CA, o AWS IoT Core oferece suporte especializado para provisionar, gerenciar e implantar certificados de PKI em dispositivos de IoT. O AWS IoT Core fornece mecanismos especializados para integração de dispositivos de IoT à infraestrutura de chave pública em grande escala.

Considerações para estabelecer uma hierarquia de CA privada

Quando precisar estabelecer uma CA privada, é importante tomar cuidado especial para projetar adequadamente a hierarquia da CA com antecedência. É uma prática recomendada implantar cada nível de sua hierarquia de CA em Contas da AWS separadas ao criar uma hierarquia de CA privada. Essa etapa intencional reduz a área de superfície de cada nível na hierarquia da CA, simplificando a descoberta de anomalias nos dados de log do CloudTrail e reduzindo o escopo de acesso ou impacto se houver acesso não autorizado a uma das contas. A CA raiz deve residir em uma própria conta separada e deve ser usada somente para emitir um ou mais certificados de CA intermediários.

Depois, crie uma ou mais CAs intermediárias em contas separadas da conta da CA raiz para emitir certificados para usuários finais, dispositivos ou outras workloads. Por fim, emita certificados da CA raiz para as CAs intermediárias, que, por sua vez, emitirão certificados para os usuários finais ou dispositivos. Para obter mais informações sobre como planejar a implantação de CA e projetar a hierarquia de CA, incluindo planejamento de resiliência, replicação entre regiões, compartilhamento de CAs na organização e muito mais, consulte Planning your AWS Private CA deployment.

Etapas da implementação

1. Determine os serviços da AWS relevantes e necessários para seu caso de uso:
    - Muitos casos de uso podem aproveitar a infraestrutura de chave pública da AWS existente usando o AWS Certificate Manager. O ACM pode ser usado para implantar certificados TLS para servidores web, balanceadores de carga ou outros usos para certificados publicamente confiáveis.

- Considere AWS Private CA quando precisar estabelecer a própria hierarquia de autoridade de certificação privada ou precisar acessar certificados exportáveis. O ACM pode então ser usado para emitir muitos tipos de certificados de entidade final usando a AWS Private CA.
- Para casos de uso em que os certificados devem ser provisionados em grande escala para dispositivos incorporados de Internet das Coisas (IoT), pense no AWS IoT Core.

2. Implemente a renovação automática do certificado sempre que possível:
   - Use a renovação gerenciada pelo ACM para certificados emitidos pelo ACM junto com serviços gerenciados da AWS integrados.
3. Estabeleça trilhas de auditoria e registro:
   - Habilite o Logs do CloudTrail para monitorar o acesso às contas que têm autoridades de certificação. Considere configurar a validação da integridade do arquivo de log no CloudTrail para verificar a autenticidade dos dados de log.
   - Gere e revise periodicamente relatórios de auditoria que listam os certificados que a CA privada emitiu ou revogou. Esses relatórios podem ser exportados para um bucket do S3.
   - Ao implantar uma CA privada, você também precisará estabelecer um bucket do S3 para armazenar a lista de revogação de certificados (CRL). Para obter orientação sobre como configurar esse bucket do S3 com base nos requisitos da workload, consulte Planejar uma lista de revogação de certificados (CRL).

Recursos

Práticas recomendadas relacionadas:

- SEC02-BP02 Usar credenciais temporárias
- SEC08-BP01 Implementar gerenciamento de chaves seguro
- SEC09-BP04 Autenticar as comunicações de rede

Documentos relacionados:

- How to host and manage an entire private certificate infrastructure in AWS
- How to secure an enterprise scale ACM Private CA hierarchy for automotive and manufacturing
- Práticas recomendadas de CA privada
- How to use AWS RAM to share your ACM Private CA cross-account

Vídeos relacionados:

- Activating AWS Certificate Manager Private CA (workshop)

Exemplos relacionados:

- Workshop de CA privada
- IOT Device Management Workshop (incluindo provisionamento de dispositivos)

Ferramentas relacionadas:

- Plug-in para o gerenciador de certificados do Kubernetes para usar a AWS Private CA

SEC09-BP02 Aplicar a criptografia em trânsito

Aplique os requisitos de criptografia definidos com base em políticas, obrigações regulatórias e padrões da organização para cumprir os requisitos organizacionais, legais e de conformidade. Utilize somente protocolos com criptografia ao transmitir dados sigilosos para fora da sua nuvem privada virtual (VPC). A criptografia ajuda a manter a confidencialidade dos dados mesmo quando os dados passam por redes não confiáveis.

Resultado desejado: todos os dados devem ser criptografados em trânsito com pacotes de criptografia e protocolos TLS seguros. O tráfego de rede entre seus recursos e a Internet deve ser criptografado para reduzir o acesso não autorizado aos dados. O tráfego de rede exclusivamente em seu ambiente interno da AWS deve ser criptografado com TLS sempre que possível. A rede interna da AWS é criptografada por padrão e o tráfego de rede em uma VPC não pode ser adulterado nem interceptado a menos que uma parte não autorizada tenha obtido acesso ao recurso que esteja gerando o tráfego (como instâncias do Amazon EC2 e contêineres do Amazon ECS). Considere proteger o tráfego de rede para rede com uma rede privada virtual (VPN) IPsec.

Antipadrões comuns:

- Utilizar versões obsoletas de SSL, TLS e componentes do pacote de criptografia (por exemplo, SSL v3.0, chaves RSA de 1024 bits e criptografia RC4).
- Permitir tráfego não criptografado (HTTP) para ou de recursos voltados para o público.
- Não monitorar e substituir certificados X.509 antes da validade.
- Utilizar certificados X.509 autoassinados para TLS.

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

Orientação de implementação

Os serviços da AWS fornecem endpoints HTTPS usando TLS para comunicação, fornecendo criptografia em trânsito quando se comunicam 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. 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 ou AWS Direct Connect para facilitar a criptografia do tráfego. Verifique se os seus clientes estão fazendo chamadas para APIs da AWS utilizando pelo menos TLS 1.2, pois a AWS tornará obsoleto o uso de TLS 1.0 e 1.1 em junho de 2023. Soluções de terceiros estão disponíveis no AWS Marketplace, caso você tenha requisitos especiais.

Etapas da implementação

- Aplicar 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 EC2.
- Configurar protocolos seguros em serviços de borda: configure o HTTPS com Amazon CloudFront e utilize um perfil de segurança apropriado para seu procedimento de segurança e caso de uso.
- Utilizar uma VPN para conectividade externa: considere usar uma VPN IPsec para proteger conexões ponto a ponto ou rede a rede para fornecer privacidade e integridade dos dados.
- Configurar protocolos seguros em balanceadores de carga: selecione uma política de segurança que ofereça os pacotes de criptografia mais fortes compatíveis com os clientes que se conectarão ao receptor. Criar um receptor de HTTPS para seu Application Load Balancer.
- Configurar protocolos seguros no Amazon Redshift: configure o cluster para exigir uma conexão Secure Socket Layer (SSL) ou Transport Layer Security (TLS).
- Configurar protocolos seguros: leia a documentação do serviço da AWS para determinar os recursos de criptografia em trânsito.
- Configurar o acesso seguro ao fazer upload para buckets do Amazon S3: utilize controles de política de bucket do Amazon S3 para implementar acesso seguro aos dados.
- Considerar o uso do AWS Certificate Manager: o ACM permite fornecer, gerenciar e implantar certificados TLS públicos para uso com serviços da AWS.

- Considerar o uso do AWS Private Certificate Authority para necessidades de PKI privada: o AWS Private CA permite criar hierarquias de autoridade de certificado privada (CA) para emitir certificados X.509 entidade final que podem ser usados para criar canais de TLS criptografados.

Recursos

Documentos relacionados:

- Documentação da AWS
- Utilizar HTTPS com o CloudFront
- Conectar sua VPC a redes remotas usando a AWS Virtual Private Network
- Criar um receptor de HTTPS para seu Application Load Balancer
- Tutorial: configurar o SSL/TLS no Amazon Linux 2
- Usar SSL/TLS para criptografar uma conexão com uma instância de banco de dados
- Configurar as opções de segurança para conexões

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.

Projete a workload para usar protocolos de rede seguros e autenticados sempre que for feita uma comunicação entre serviços, aplicações ou usuários. O uso de protocolos de rede compatíveis com a autenticação e a autorização fornece maior controle sobre os fluxos de rede e reduz o impacto do acesso não autorizado.

Resultado desejado: uma workload com fluxos de tráfego bem definidos do plano de dados e do ambiente de gerenciamento entre os serviços. Os fluxos de tráfego usam protocolos de rede autenticados e criptografados quando tecnicamente viáveis.

Antipadrões comuns:

- Fluxos de tráfego não criptografados ou não autenticados na workload.
- Reutilizar credenciais de autenticação entre vários usuários ou entidades.
- Confiar apenas nos controles de rede como um mecanismo de controle de acesso.
- Criar um mecanismo de autenticação personalizado em vez de depender de mecanismos de autenticação padrão do setor.
- Fluxos de tráfego excessivamente permissivos entre componentes de serviço ou outros recursos na VPC.

Benefícios do estabelecimento desta prática recomendada:

- Limita o escopo do impacto do acesso não autorizado a uma parte da workload.
- Fornece um nível mais alto de garantia de que as ações são executadas somente por entidades autenticadas.
- Melhora o desacoplamento de serviços definindo e aplicando claramente as interfaces de transferência de dados pretendidas.
- Melhora o monitoramento, o log e a resposta a incidentes por meio da atribuição de solicitações e interfaces de comunicação bem definidas.
- Oferece defesa profunda para as workloads combinando controles de rede com controles de autenticação e de autorização.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: baixo

Orientações para a implementação

Os padrões de tráfego de rede da workload podem ser caracterizados em duas categorias:

- O tráfego leste-oeste representa fluxos de tráfego entre serviços que compõem uma workload.
- O tráfego norte-sul representa fluxos de tráfego entre a workload e os consumidores.

Embora seja uma prática comum criptografar o tráfego norte-sul, é menos comum proteger o tráfego leste-oeste usando protocolos autenticados. As práticas modernas de segurança recomendam que o design da rede por si só não conceda um relacionamento confiável entre duas entidades. Quando dois serviços puderem residir dentro de um limite de rede comum, criptografar, autenticar e autorizar as comunicações ainda são práticas recomendadas entre esses serviços.

Como exemplo, as APIs de serviços da AWS usam o protocolo de assinatura do Signature Version 4 (SigV4) da AWS para autenticar o chamador, independentemente da rede de origem da solicitação. Essa autenticação garante que as APIs da AWS possam verificar a identidade que solicitou a ação e que essa identidade possa ser combinada com políticas para tomar uma decisão de autorização a fim de determinar se a ação deve ser permitida ou não.

Serviços, como o Amazon VPC Lattice e o Amazon API Gateway permitem usar o mesmo protocolo de assinatura SigV4 para adicionar autenticação e autorização ao tráfego leste-oeste em suas próprias workloads. Se os recursos fora do ambiente da AWS precisarem se comunicar com os serviços que exigem autenticação e autorização baseadas em SigV4, você poderá usar o AWS Identity and Access Management (IAM) Roles Anywhere no recurso que não é da AWS para adquirir

credenciais temporárias da AWS. Essas credenciais podem ser usadas para assinar solicitações para serviços que usam o SigV4 para autorizar o acesso.

Outro mecanismo comum para autenticar o tráfego leste-oeste é a autenticação mútua TLS (mTLS). Muitas aplicações da Internet das Coisas (IoT), aplicações business to business e microsserviços usam o mTLS para validar a identidade de ambos os lados de uma comunicação TLS por meio do uso de certificados X.509 do lado do cliente e do servidor. Esses certificados podem ser emitidos por AWS Private Certificate Authority (AWS Private CA). É possível usar serviços como o Amazon API Gateway e o AWS App Mesh para fornecer autenticação mTLS para comunicação entre workloads ou dentro da workload. Embora o mTLS forneça informações de autenticação aos dois lados de uma comunicação TLS, ele não fornece um mecanismo de autorização.

Por fim, o OAuth 2.0 e o OpenID Connect (OIDC) são dois protocolos normalmente usados para controlar o acesso aos serviços pelos usuários, mas agora também estão se tornando populares para o tráfego entre serviços. O API Gateway fornece um autorizador JSON Web Token (JWT), que permite que as workloads restrinjam o acesso às rotas de API usando JWTs emitidos por provedores de identidades OIDC ou OAuth 2.0. Os escopos do OAuth2 podem ser usados como uma fonte para decisões básicas de autorização, mas as verificações de autorização ainda precisam ser implementadas na camada da aplicação, e os escopos do OAuth2 por si sós não atendem a necessidades de autorização mais complexas.

Etapas da implementação

- Definir e documentar os fluxos de rede da workload: a primeira etapa na implementação de uma estratégia de defesa profunda é definir os fluxos de tráfego da workload.
  - Crie um diagrama de fluxo de dados que defina claramente como os dados são transmitidos entre os diferentes serviços que compõem a workload. Esse diagrama é a primeira etapa para aplicar esses fluxos por meio de canais de rede autenticados.
  - Instrumente a workload nas fases de desenvolvimento e testes para validar se o diagrama de fluxo de dados reflete com precisão o comportamento da workload em tempo de execução.
  - Um diagrama de fluxo de dados também pode ser útil ao realizar um exercício de modelagem de ameaças, conforme descrito em SEC01-BP07 Identificar ameaças e priorizar mitigações com o uso de um modelo de ameaça.
- Estabeleça controles de rede: considere os recursos da AWS para estabelecer controles de rede alinhados aos fluxos de dados. Embora os limites da rede não devam ser o único controle de segurança, eles fornecem uma camada na estratégia de defesa profunda para proteger a workload.

  - Use grupos de segurança para estabelecer, definir e restringir fluxos de dados entre recursos.
  - Considere usar o AWS PrivateLink para se comunicar com os serviços da AWS e de terceiros que são compatíveis com o AWS PrivateLink. Os dados enviados por meio de um endpoint da interface do AWS PrivateLink permanecem na estrutura da rede da AWS e não atravessam a internet pública.
- Implementar autenticação e autorização entre os serviços na workload: escolha o conjunto de serviços da AWS mais apropriado para fornecer fluxos de tráfego autenticados e criptografados na workload.
  - Considere o Amazon VPC Lattice para proteger a comunicação entre serviços. O VPC Lattice pode usar a autenticação do SigV4 combinada com políticas de autenticação para controlar o acesso entre serviços.
  - Para comunicação entre serviços usando mTLS, considere o API Gateway ou o App Mesh. O AWS Private CA pode ser usado para estabelecer uma hierarquia de CA privada capaz de emitir certificados para uso com o mTLS.
  - Ao fazer a integração com serviços que usam OAuth 2.0 ou OIDC, considere o API Gateway usando o autorizador JWT.
  - Para comunicação entre a workload e dispositivos de IoT, considere o AWS IoT Core, que fornece várias opções para criptografia e autenticação de tráfego de rede.
- Monitorar o acesso não autorizado: monitore continuamente os canais de comunicação não intencionais, entidades principais não autorizadas que tentam acessar recursos protegidos e outros padrões de acesso inadequados.
  - Se estiver usando o VPC Lattice para gerenciar o acesso aos serviços, considere ativar e monitorar os logs de acesso do VPC Lattice. Esses logs de acesso incluem informações sobre a entidade solicitante, informações de rede que incluem a VPC de origem e de destino e os metadados da solicitação.
  - Considere a ativação dos Logs de fluxo da VPC para capturar metadados nos fluxos de rede e analisar se há anomalias periodicamente.
  - Consulte o AWS Security Incident Response Guide e a seção Resposta a incidentes do Pilar Segurança: AWS Well-Architected Framework para obter mais orientações sobre planejamento, simulação e resposta a incidentes de segurança.

Recursos

Práticas recomendadas relacionadas:

- SEC03-BP07 Analisar o acesso público e entre contas
- SEC02-BP02 Usar credenciais temporárias
- SEC01-BP07 Identificar ameaças e priorizar mitigações com o uso de um modelo de ameaça

Documentos relacionados:

- Evaluating access control methods to secure Amazon API Gateway APIs
- Configurar a autenticação TLS mútua para uma API REST
- How to secure API Gateway HTTP endpoints with JWT authorizer
- Authorizing direct calls to AWS services using AWS IoT Core credential provider
- Guia de resposta a incidentes de segurança da AWS

Vídeos relacionados:

- AWS re:invent 2022: Introducing VPC Lattice
- AWS re:invent 2020: Serverless API authentication for HTTP APIs on AWS

Exemplos relacionados:

- Workshop do Amazon VPC Lattice
- Workshop Zero-Trust Episode 1 – The Phantom Service Perimeter

Resposta a incidentes

Pergunta

- SEGURANÇA 10. Como prever, responder e se recuperar de incidentes?

SEGURANÇA 10. Como prever, responder e se recuperar de incidentes?

Mesmo com controles preventivos e de detecção consolidados, sua organização deve implementar processos para responder e reduzir o impacto potencial de incidentes de segurança. Sua preparação afeta muito a capacidade das equipes operarem efetivamente durante um incidente, isolarem, conterem e analisarem problemas e restaurarem as operações para um estado adequado conhecido. Implementar as ferramentas e o acesso antes de um incidente de segurança e praticar

rotineiramente dias de jogos para validar a resposta a incidentes ajuda a garantir que você possa se recuperar enquanto minimiza interrupções empresariais.

Práticas recomendadas

- SEC10-BP01 Identify key personnel and external resources
- SEC10-BP02 Desenvolver planos de gerenciamento de incidentes
- SEC10-BP03 Prepare recursos forenses
- SEC10-BP04 Desenvolva e teste manuais de resposta a incidentes de segurança
- SEC10-BP05 Acesso pré-provisionado
- SEC10-BP06 Pré-implantação de ferramentas
- SEC10-BP07 Execute simulações
- SEC10-BP08 Estabeleça uma estrutura para aprender com os incidentes

SEC10-BP01 Identify key personnel and external resources

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 para implementação

- Identificar o pessoal-chave da organização: Mantenha uma lista de contatos da sua organização que você precisaria acionar para responder e recuperar-se de um incidente.

- Identificar parceiros externos: Entre em contato com parceiros externos, se necessário, que possam ajudá-lo 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

O primeiro documento a ser desenvolvido para resposta a incidentes é o plano de resposta a incidentes. O plano de resposta a incidentes foi projetado para ser a base de seu programa e estratégia de resposta a incidentes.

Benefícios de estabelecer esta prática recomendada: O desenvolvimento de processos de resposta a incidentes completos e claramente definidos é fundamental para um programa de resposta a incidentes bem-sucedido e escalável. Quando ocorre um evento de segurança, etapas e fluxos de trabalho claros poderão ajudar você a responder em tempo hábil. Talvez você já tenha processos de resposta a incidentes existentes. Independentemente do seu estado atual, é importante atualizar, repetir e testar seus processos de resposta a incidentes regularmente.

Nível de risco exposto se esta 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 informações de identificação pessoal (PII) da Europa, considere cenários como a maneira 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 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.

Etapas da implementação

Defina funções e responsabilidades

Lidar com eventos de segurança exige disciplina interorganizacional e uma inclinação para a ação. Em sua estrutura organizacional, deve haver muitas pessoas responsáveis, atribuídas, consultadas ou mantidas informadas durante um incidente, como representantes de recursos humanos (RH), da equipe executiva e do setor jurídico. Considere essas funções e responsabilidades e se algum terceiro deve estar envolvido. Observe que muitas regiões têm leis locais que regem o que deve e o que não deve ser feito. Embora possa parecer burocrático criar um grafo de pessoas responsáveis, atribuídas, consultadas e informadas (RACI) para seus planos de resposta de segurança, isso facilita a comunicação rápida e direta e descreve claramente a liderança em diferentes estágios do evento.

Durante um incidente, incluir os proprietários e os desenvolvedores de aplicações e recursos afetados é fundamental porque eles são especialistas no assunto (PMEs) que podem fornecer informações e contexto para ajudar a medir o impacto. Pratique e construa relacionamentos com os desenvolvedores e os proprietários de aplicações antes de confiar na experiência deles para responder a incidentes. Proprietários de aplicações ou PMEs, como administradores ou engenheiros

de nuvem, podem precisar agir em situações em que o ambiente não seja familiar ou tenha complexidade, ou em que os respondentes não tenham acesso.

Por fim, parceiros confiáveis podem estar envolvidos na investigação ou na resposta, pois podem oferecer experiência adicional e um controle valioso. Se você não tiver essas habilidades em sua própria equipe, contrate uma parte externa para obter assistência.

Entender as equipes de resposta e o suporte da AWS

- AWS Support
  - O AWS Support oferece uma variedade de planos que concedem acesso a ferramentas e conhecimentos que apoiam o êxito e a saúde operacional de suas soluções da AWS. Se precisar de suporte técnico e mais recursos para ajudar a planejar, implantar e otimizar seu ambiente da AWS, selecione um plano de suporte mais adequado ao seu caso de uso da AWS.
  - Considere o Support Center entre AWS Management Console (é necessário fazer login) como ponto central de contato para obter suporte para problemas que afetam seus recursos da AWS. O acesso ao AWS Support é controlado pelo AWS Identity and Access Management. Para ter mais informações sobre como obter acesso aos recursos da AWS Support, consulte Conceitos básicos do AWS Support.
- Equipe de Resposta a Incidentes de Clientes (CIRT) da AWS
  - A Equipe de Resposta a Incidentes de Clientes (CIRT) da AWS é uma equipe global da AWS especializada 24 horas por dia, 7 dias por semana, que presta assistência aos clientes durante eventos de segurança ativos do cliente do Modelo de responsabilidade compartilhada da AWS.
  - Ao apoiar você, a AWS CIRT presta assistência na triagem e na recuperação de um evento de segurança ativo na AWS. Eles podem ajudar na análise da causa raiz por meio do uso de logs de serviço da AWS e fornecer recomendações para recuperação. Eles também podem fornecer recomendações de segurança e práticas recomendadas para ajudar você a evitar eventos de segurança no futuro.
  - Os clientes da AWS podem contratar a AWS CIRT por meio de um caso do AWS Support.
- Suporte de resposta a DDoS
  - A AWS oferece o AWS Shield, que fornece um serviço gerenciado de proteção distribuída de negação de serviço (DDoS) que protege as aplicações web em execução na AWS. O Shield oferece detecção contínua e mitigações automáticas em linha que podem minimizar o tempo de inatividade e a latência da aplicação, portanto, não há necessidade de contratar o AWS Support para se beneficiar da proteção contra DDoS. Existem dois níveis de Shield: AWS Shield

Standard e AWS Shield Advanced. Para saber mais sobre as diferenças entre esses dois níveis, consulte a documentação de recursos do Shield.

- AWS Managed Services (AMS)
  - O AWS Managed Services (AMS) oferece gerenciamento contínuo de sua infraestrutura da AWS para que você possa se concentrar em suas aplicações. Ao implementar as práticas recomendadas para manter sua infraestrutura, o AMS ajuda a reduzir sua sobrecarga operacional e os riscos. O AMS automatiza atividades comuns, como solicitações de mudança, monitoramento, gerenciamento de patches, serviços de segurança e backup, e fornece serviços de ciclo de vida completo para provisionar, executar e oferecer compatibilidade com sua infraestrutura.
  - O AMS assume a responsabilidade de implantar um pacote de controles de detecção de segurança e fornece uma primeira linha de resposta aos alertas 24 horas por dia, 7 dias por semana. Quando um alerta é iniciado, o AMS segue um conjunto padrão de guias e manuais automatizados para verificar uma resposta consistente. Esses guias são compartilhados com os clientes do AMS durante a integração para que eles possam desenvolver e coordenar uma resposta com o AMS.

Desenvolva o plano de resposta a incidentes

O plano de resposta a incidentes foi projetado para ser a base de seu programa e estratégia de resposta a incidentes. O plano de resposta a incidentes deve estar em um documento formal. Um plano de resposta a incidentes geralmente inclui as seguintes seções:

- Uma visão geral da equipe de resposta a incidentes: Descreve as metas e as funções da equipe de resposta a incidentes.
- Funções e responsabilidades: Lista as partes interessadas na resposta a incidentes e detalha suas funções quando ocorre um incidente.
- Um plano de comunicação: Detalha as informações de contato e como você se comunica durante um incidente.
- Métodos de comunicação de backup: É prática recomendada ter a comunicação fora de banda como backup para a comunicação de incidentes. Um exemplo de aplicação que fornece um canal seguro de comunicação fora de banda é AWS Wickr.
- Fases da resposta a incidentes e ações a serem realizadas: Enumera as fases da resposta a incidentes (por exemplo, detectar, analisar, erradicar, conter e recuperar), incluindo ações de alto nível a serem realizadas nessas fases.

- Definições de severidade e priorização do incidente: Detalha como classificar a severidade de um incidente, como priorizar o incidente e, depois, como as definições de severidade afetam os procedimentos de escalonamento.

Embora essas seções sejam comuns em empresas de diferentes tamanhos e setores, o plano de resposta a incidentes de cada organização é único. Você precisa criar um plano de resposta a incidentes que funcione melhor para a organização.

Recursos

Práticas recomendadas relacionadas:

- SEC04 (Como você detecta e investiga eventos de segurança?)

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

SEC10-BP03 Prepare recursos forenses

Antes de um incidente de segurança, considere o desenvolvimento de recursos forenses para apoiar as investigações de eventos de segurança.

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

Os conceitos da análise forense on-premises tradicional se aplicam à AWS. Para obter informações importantes para começar a desenvolver recursos forenses na Nuvem AWS, consulte Forensic investigation environment strategies in the Nuvem AWS.

Depois de configurar o ambiente e a estrutura da Conta da AWS para análise forense, defina as tecnologias necessárias para executar com eficácia metodologias forenses sólidas nas quatro fases:

- Coleta: Colete logs relevantes da AWS, como logs do AWS CloudTrail, do AWS Config, logs de fluxo da VPC e log em nível de host. Colete snapshots, backups e despejos de memória dos recursos afetados da AWS, quando disponíveis.
- Exame: Examine os dados coletados extraindo e avaliando as informações relevantes.
- Análises: Analise os dados coletados para entender o incidente e tirar conclusões dele.

- Relatórios: Apresente as informações resultantes da fase de análise.

Etapas da implementação

Prepare o ambiente forense

AWS Organizations ajuda a gerenciar e reger centralmente um ambiente da AWS à medida que você expande e escala os recursos da AWS. Uma organização da AWS consolida suas Contas da AWS para que você possa administrá-las como uma única unidade. Você pode usar unidades organizacionais (UOs) para agrupar contas e administrá-las como uma única unidade.

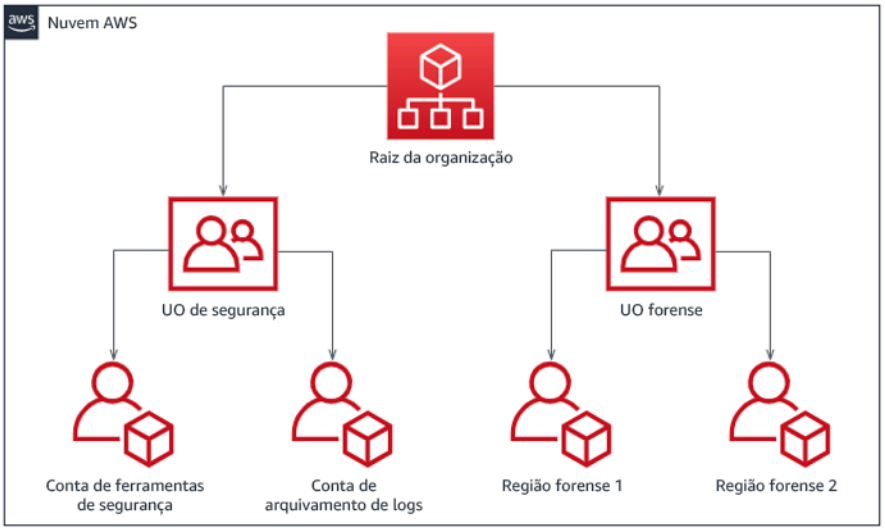
Para resposta a incidentes, é útil ter uma estrutura da Conta da AWS compatível com as funções de resposta a incidentes, que inclui uma UO de segurança e uma UO forense. Dentro da OU de segurança, você deve ter contas para:

- Arquivamento de logs: Agregue logs em uma Conta da AWS de arquivamento de logs com permissões limitadas.
- Ferramentas de segurança: Centralize os serviços de segurança em uma Conta da AWS de ferramenta de segurança. Essa conta opera como administrador delegado dos serviços de segurança.

Dentro da UO forense, você tem a opção de implementar uma única conta ou contas forenses para cada região em que opera, dependendo da que funciona melhor para sua empresa e modelo operacional. Se você criar uma conta forense por região, poderá bloquear a criação de recursos da AWS fora dessa região e reduzir o risco de os recursos serem copiados para uma região não pretendida. Por exemplo, se você opera apenas em US East (N. Virginia) Region (`us-east-1`) e US West (Oregon) (`us-west-2`), então você teria duas contas na UO forense: uma para `us-east-1` e uma para `us-west-2`.

Você pode criar uma Conta da AWS de análise forense para várias regiões. Você deve ter cuidado ao copiar recursos da AWS para essa conta para verificar se está de acordo com seus requisitos de soberania de dados. Como é preciso tempo para provisionar novas contas, é imperativo criar e instrumentar as contas forenses bem antes de um incidente, para que os respondentes possam estar preparados para usá-las de forma eficaz em suas respostas.

O diagrama a seguir exibe um exemplo de estrutura de contas, incluindo uma UO forense com contas forenses por região:



Estrutura de contas por região para resposta a incidentes

Capture backups e snapshots

Configurar backups dos principais sistemas e bancos de dados é essencial para a recuperação de um incidente de segurança e para fins forenses. Com os backups em vigor, você pode restaurar seus sistemas ao estado seguro anterior. Na AWS, você pode criar snapshots de vários recursos. Os snapshots fornecem backups pontuais desses recursos. Há muitos serviços da AWS que podem ajudar em backup e recuperação. Para obter detalhes sobre esses serviços e abordagens para backup e recuperação, consulte Backup and Recovery Prescriptive Guidance e Use backups to recover from security incidents.

Especialmente quando se trata de situações como ransomware, é fundamental que os backups estejam bem protegidos. Para obter orientações sobre como proteger os backups, consulte Top 10 security best practices for securing backups in AWS. Além de proteger os backups, você deve testar regularmente seus processos de backup e restauração para verificar se a tecnologia e os processos implementados funcionam conforme o esperado.

Automatize a análise forense

Durante um evento de segurança, sua equipe de resposta a incidentes deve ser capaz de coletar e analisar evidências rapidamente, mantendo a precisão durante o período em torno do evento (como capturar registros relacionados a um evento ou recurso específico ou coletar o despejo de memória de uma instância do Amazon EC2). É desafiador e demorado para a equipe de resposta a incidentes coletar manualmente as evidências relevantes, especialmente em um grande número de instâncias

e contas. Além disso, a coleta manual pode estar sujeita a erros humanos. Por esses motivos, você deve desenvolver e implementar a automação para perícia o máximo possível.

A AWS oferece vários recursos de automação para análise forense, que estão listados na seção de recursos a seguir. Esses recursos são exemplos de padrões forenses que desenvolvemos e que os clientes implementaram. Embora possam ser uma arquitetura de referência útil para começar, considere modificá-las ou criar padrões de automação forense com base em seu ambiente, requisitos, ferramentas e processos forenses.

Recursos

Documentos relacionados:

- AWS Security Incident Response Guide - Develop Forensics Capabilities
- AWS Security Incident Response Guide - Forensics Resources
- Forensic investigation environment strategies in the Nuvem AWS
- How to automate forensic disk collection in AWS
- AWS Prescriptive Guidance - Automate incident response and forensics

Vídeos relacionados:

- Automatização de resposta a incidentes e forense

Exemplos relacionados:

- Automated Incident Response and Forensics Framework (Estrutura forense e de resposta automatizada a incidentes)
- Automated Forensics Orchestrator for Amazon EC2

SEC10-BP04 Desenvolva e teste manuais de resposta a incidentes de segurança

Uma parte fundamental da preparação de seus processos de resposta a incidentes é desenvolver manuais. Os manuais de resposta a incidentes fornecem uma série de orientações prescritivas e etapas a serem seguidas quando ocorre um evento de segurança. Ter uma estrutura e etapas claras simplifica a resposta e reduz a probabilidade de erro humano.

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

Orientação para implementação

Os manuais devem ser criados para cenários de incidentes, como:

- Incidentes esperados: os manuais devem ser criados para os incidentes previstos. Isso inclui ameaças como negação de serviço (DoS), ransomware e comprometimento de credenciais.
- Descobertas ou alertas de segurança conhecidos: os manuais devem ser criados para descobertas e alertas de segurança conhecidos, como descobertas do GuardDuty. Você pode receber uma descoberta do GuardDuty e pensar: “E agora?” Para evitar que você trate incorretamente ou ignore uma descoberta do GuardDuty, crie um manual para cada possível descoberta do GuardDuty. Alguns detalhes e orientações sobre a correção podem ser encontrados na documentação do GuardDuty. É importante notar que o GuardDuty não está habilitado por padrão e tem um custo. Para obter mais detalhes sobre o GuardDuty, consulte Appendix A: Cloud capability definitions - Visibility and alerting.

Os manuais devem conter etapas técnicas a serem concluídas por um analista de segurança para investigar e responder adequadamente a um possível incidente de segurança.

Etapas da implementação

Os itens a serem incluídos em um manual incluem:

- Visão geral do manual: qual cenário de risco ou incidente esse manual aborda? Qual é o objetivo do manual?
- Pré-requisitos: quais logs, mecanismos de detecção e ferramentas automatizadas são necessários para esse cenário de incidente? Qual é a notificação esperada?
- Informações de comunicação e escalonamento: quem está envolvido e quais são suas informações de contato? Quais são as responsabilidades de cada parte interessada?
- Etapas de resposta: em todas as fases da resposta a incidentes, quais etapas táticas devem ser seguidas? Quais consultas um analista deve executar? Qual código deve ser executado para alcançar o resultado desejado?
  - Detectar: como o incidente será detectado?
  - Análise: como o escopo do impacto será determinado?
  - Contêm: como o incidente será isolado para limitar o escopo?
  - Erradicar: como a ameaça será removida do ambiente?
  - Recuperar: como o sistema ou o recurso afetado voltará à produção?

- Resultados esperados: depois que as consultas e o código forem executados, qual é o resultado esperado do manual?

Recursos

Práticas recomendadas relacionadas ao Well-Architected:

- SEC10-BP02 - Develop incident management plans (SEC10-BP02 – Desenvolver planos de gerenciamento de incidentes)

Documentos relacionados:

- Framework for Incident Response Playbooks (Estrutura para manuais de resposta a incidentes)
- Develop your own Incident Response Playbooks (Desenvolva seus próprios manuais de resposta a incidentes)
- Incident Response Playbook Samples (Amostras do manual de resposta a incidentes)
- Building an AWS incident response runbook using Jupyter playbooks and CloudTrail Lake

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 Pré-implantação de ferramentas

Verifique se o pessoal de segurança tem as ferramentas certas pré-implantadas para reduzir o tempo de investigação até a recuperação.

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

Orientação para implementação

Para automatizar as funções de resposta 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.

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.

Durante uma investigação de segurança, você precisa ser capaz de analisar os logs relevantes para registrar e compreender o escopo completo e o cronograma do incidente. Os logs também são necessários para geração de alertas indicando que ocorreram determinadas ações de interesse. É essencial selecionar, ativar, armazenar e configurar mecanismos de consulta, recuperação e definir

alertas. Além disso, uma forma eficaz de fornecer ferramentas para pesquisar dados de log é o Amazon Detective.

A AWS oferece mais de 200 serviços em nuvem e milhares de recursos. Recomendamos que você analise os serviços que podem apoiar e simplificar sua estratégia de resposta a incidentes.

Além do registro em log, você deve desenvolver e implementar uma estratégia consistente de marcação. A marcação pode ajudar a fornecer contexto sobre a finalidade de um recurso da AWS. A marcação também pode ser usada para automação.

Etapas da implementação

Selecione e configure logs para análise e alertas

Consulte a documentação a seguir sobre como configurar logs para resposta a incidentes:

- Logging strategies for security incident response (Estratégias de registro para resposta a incidentes de segurança)
- SEC04-BP01 Configurar registro em log de serviço e aplicação

Habilite serviços de segurança para oferecer suporte à detecção e resposta

A AWS fornece recursos nativos de detecção, prevenção e resposta, e outros serviços podem ser usados para arquitetar soluções de segurança personalizadas. Para obter uma lista dos serviços mais relevantes para resposta a incidentes de segurança, consulte Definições de capacidade de nuvem.

Desenvolva e implemente uma estratégia de marcação

Obter informações contextuais sobre o caso de uso empresarial e as partes interessadas internas relevantes em torno de um recurso da AWS pode ser difícil. Uma forma de fazer isso é na forma de tags, que atribuem metadados aos recursos da AWS e consistem em uma chave e um valor definidos pelo usuário. Você pode criar tags para categorizar os recursos por finalidade, proprietário, ambiente, tipo de dados processados e outros critérios de sua escolha.

Ter uma estratégia de marcação consistente pode acelerar os tempos de resposta e minimizar o tempo gasto no contexto organizacional, permitindo identificar e discernir rapidamente as informações contextuais sobre um recurso da AWS. As tags também podem servir como um mecanismo para iniciar automações de resposta. Para obter mais detalhes sobre o que marcar,

consulte Tagging your AWS resources. Primeiro, você deve definir as tags que deseja implementar em toda a sua organização. Depois disso, você implementará e aplicará essa estratégia de marcação. Para obter mais detalhes sobre implementação e aplicação, consulte Implement AWS resource tagging strategy using AWS Tag Policies and Service Control Policies (SCPs).

Recursos

Práticas recomendadas relacionadas ao Well-Architected:

- SEC04-BP01 Configurar registro em log de serviço e aplicação
- SEC04-BP02 Analisar logs, descobertas e métricas de forma centralizada

Documentos relacionados:

- Logging strategies for security incident response (Estratégias de registro para resposta a incidentes de segurança)
- Incident response cloud capability definitions (Definições de recursos de nuvem de resposta a incidentes)

Exemplos relacionados:

- Threat Detection and Response with Amazon GuardDuty and Amazon Detective
- Security Hub Workshop (Workshop do Security Hub)
- Vulnerability Management with Amazon Inspector

SEC10-BP07 Execute simulações

À medida que as organizações crescem e evoluem com o tempo, o mesmo acontece com o cenário de ameaças, o que torna importante analisar continuamente seus recursos de resposta a incidentes. Executar simulações (também conhecidas como dias de teste) é um método que pode ser usado para realizar essa avaliação. As simulações usam cenários de eventos de segurança do mundo real projetados para imitar as táticas, as técnicas e os procedimentos (TTPs) de um agente de ameaças e permitir que uma organização exercite e avalie seus recursos de resposta a incidentes respondendo a esses eventos cibernéticos simulados da mesma forma que em uma situação real.

Benefícios do estabelecimento dessa prática recomendada: as simulações têm vários benefícios:

- Validar a prontidão cibernética e desenvolver a confiança de seus socorristas.

- Testar a precisão e a eficiência de ferramentas e fluxos de trabalho.
- Refinar os métodos de comunicação e escalonamento alinhados com seu plano de resposta a incidentes.
- Proporcionar uma oportunidade de responder a vetores menos comuns.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

Existem três tipos principais de simulações:

- Simulações teóricas: a abordagem de simulações teóricas é uma sessão baseada em discussões que envolvem as várias partes interessadas na resposta a incidentes para exercer funções e responsabilidades e usar ferramentas de comunicação e manuais estabelecidos. A facilitação das simulações normalmente pode ser realizada em um dia inteiro em um local virtual, local físico ou uma combinação de ambos. Por ser baseada em discussões, a simulação teórica se concentra em processos, pessoas e colaboração. A tecnologia é parte integrante da discussão, mas o uso real de ferramentas ou scripts de resposta a incidentes geralmente não faz parte da simulação teórica.
- Simulações da equipe roxa: as simulações da equipe roxa aumentam o nível de colaboração entre os respondentes ao incidente (equipe azul) e os agentes de ameaças simuladas (equipe vermelha). A equipe azul é composta por membros do centro de operações de segurança (SOC), mas também pode incluir outras partes interessadas que estariam envolvidas durante um evento cibernético real. A equipe vermelha é composta por uma equipe de testes de penetração ou pelas principais partes interessadas treinadas em segurança ofensiva. A equipe vermelha trabalha em colaboração com os facilitadores da simulação ao projetar um cenário para que este seja preciso e viável. Durante as simulações da equipe roxa, o foco principal está nos mecanismos de detecção, nas ferramentas e nos procedimentos operacionais padrão (SOPs) que apoiam os esforços de resposta a incidentes.
- Simulações da equipe vermelha: durante uma simulação da equipe vermelha, o infrator (equipe vermelha) realiza uma simulação para atingir um determinado objetivo ou conjunto de objetivos a partir de um escopo predeterminado. Os defensores (equipe azul) não necessariamente terão conhecimento do escopo e da duração da simulação, o que oferece uma avaliação mais realista de como eles responderiam a um incidente real. Como as simulações da equipe vermelha podem ser testes invasivos, tenha cuidado e implemente controles para verificar se a simulação não causa danos reais ao ambiente.

Considere facilitar as simulações cibernéticas em intervalos regulares. Cada tipo de simulação pode oferecer benefícios exclusivos aos participantes e à organização como um todo. Portanto, você pode optar por começar com tipos de simulação menos complexos (como simulações teóricas) e avançar para tipos de simulação mais complexos (simulações da equipe vermelha). Você deve selecionar um tipo de simulação com base em sua maturidade de segurança, recursos e resultados desejados. Alguns clientes podem não optar por realizar simulações da equipe vermelha devido à complexidade e ao custo.

Etapas da implementação

Independentemente do tipo de simulação que você escolher, as simulações geralmente seguem estas etapas de implementação:

1. Defina os principais elementos do exercício: defina o cenário e os objetivos da simulação. Ambos devem ter aceitação da liderança.
2. Identifique as principais partes interessadas: no mínimo, um exercício precisa de facilitadores e participantes. Dependendo do cenário, outras partes interessadas, como departamento jurídico, de comunicação ou liderança executiva, podem estar envolvidos.
3. Crie e teste o cenário: talvez o cenário precise ser redefinido durante a criação se elementos específicos não forem viáveis. Espera-se um cenário finalizado como resultado dessa etapa.
4. Facilite a simulação: o tipo de simulação determina a facilitação usada (um cenário impresso em comparação a um cenário simulado altamente técnico). Os facilitadores devem alinhar suas táticas de facilitação aos objetos da simulação e envolver todos os participantes sempre que possível para proporcionar o máximo benefício.
5. Desenvolva o relatório pós-ação (AAR): identifique as áreas que funcionaram bem, aquelas que podem ser melhoradas e possíveis déficits. O AAR deve medir a eficácia da simulação, bem como a resposta da equipe ao evento simulado, para que o progresso possa ser monitorado ao longo do tempo com simulações futuras.

Recursos

Documentos relacionados:

- AWS Incident Response Guide (Guia de resposta a incidentes da AWS)

Vídeos relacionados:

- AWS GameDay - Security Edition (Dia de jogo da AWS: edição de segurança)

SEC10-BP08 Estabeleça uma estrutura para aprender com os incidentes

A implementação de uma framework de lições aprendidas e da capacidade de análise da causa raiz não só ajudará a melhorar os recursos de resposta a incidentes, mas também ajudará a evitar que o incidente se repita. Ao aprender com cada incidente, você pode ajudar a evitar a repetição dos mesmos erros, exposições ou configurações incorretas, não apenas melhorando seu procedimento de segurança, mas também minimizando o tempo perdido em situações evitáveis.

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

Orientação para implementação

É importante implementar uma framework de lições aprendidas que estabeleça e alcance, em nível geral, os seguintes pontos:

- Quando as lições são aprendidas?
- O que está envolvido no processo de lições aprendidas?
- Como as lições aprendidas são colocadas em prática?
- Quem está envolvido no processo e como?
- Como as áreas de melhoria serão identificadas?
- Como você garantirá que as melhorias sejam monitoradas e implementadas de forma eficaz?

A estrutura não deve se concentrar em culpar os indivíduos, mas sim na melhoria de ferramentas e processos.

Etapas da implementação

Além dos resultados de alto nível listados acima, é importante garantir que você faça as perguntas certas para obter o máximo valor (informações que levem a melhorias práticas) do processo. Considere estas perguntas para ajudar você a começar a promover discussões sobre as lições aprendidas:

- Qual foi o incidente?
- Quando o incidente foi identificado pela primeira vez?
- Como ele foi identificado?
- Quais sistemas alertaram sobre a atividade?

- Quais sistemas, serviços e dados estavam envolvidos?
- O que ocorreu especificamente?
- O que funcionou bem?
- O que não funcionou bem?
- Quais processos ou procedimentos falharam ou não tiveram a escala ajustada para responder ao incidente?
- O que pode ser melhorado nas seguintes áreas:
  - Pessoas
    - As pessoas que precisavam ser contatadas estavam realmente disponíveis e a lista de contatos estava atualizada?
    - As pessoas estavam perdendo treinamentos ou não tinham os recursos necessários para responder e investigar o incidente com eficácia?
    - Os recursos apropriados estavam prontos e disponíveis?
  - Processo
    - Os processos e procedimentos foram seguidos?
    - Os processos e procedimentos foram documentados e estavam disponíveis para esse (tipo de) incidente?
    - Estavam faltando processos e procedimentos necessários?
    - Os respondentes conseguiram obter acesso oportuno às informações necessárias para responder ao problema?
  - Tecnologia
    - Os sistemas de alerta existentes identificaram e alertaram efetivamente sobre a atividade?
    - Como poderíamos ter reduzido o tempo de detecção em 50%?
    - Os alertas existentes precisam ser aprimorados ou novos alertas precisam ser criados para esse (tipo de) incidente?
    - As ferramentas existentes permitiram uma investigação (pesquisa/análise) eficaz do incidente?
    - O que pode ser feito para ajudar a identificar esse (tipo de) incidente mais cedo?
    - O que pode ser feito para ajudar a evitar que esse (tipo de) incidente ocorra novamente?
    - Quem é o proprietário do plano de melhoria e como você testará se ele foi implementado?
    - Qual é o cronograma para que os controles e processos adicionais de monitoramento ou prevenção sejam implementados e testados?

Essa lista não inclui tudo, mas serve como ponto de partida para identificar quais são as necessidades da organização e da empresa e como você pode analisá-las para aprender com os incidentes de forma mais eficaz e melhorar constantemente seu procedimento de segurança. O mais importante é começar incorporando as lições aprendidas como parte padrão do processo de resposta a incidentes, da documentação e das expectativas das partes interessadas.

Recursos

Documentos relacionados:

- AWS Security Incident Response Guide - Establish a framework for learning from incidents
- NCSC CAF guidance - Lessons learned (Orientações do NCSC CAF: lições aprendidas)

Segurança de aplicações

Pergunta

- SEGURANÇA 11. Como incorporar e validar as propriedades de segurança de aplicações durante o ciclo de vida de design, desenvolvimento e implantação?

SEGURANÇA 11. Como incorporar e validar as propriedades de segurança de aplicações durante o ciclo de vida de design, desenvolvimento e implantação?

Treinar a equipe, testar por meio da automação, entender as dependências e validar as propriedades de segurança de ferramentas e aplicações ajuda a diminuir a probabilidade de problemas de segurança em workloads de produção.

Práticas recomendadas

- SEC11-BP01 Treinar para segurança de aplicações
- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento
- SEC11-BP03 Realizar teste de penetração regular
- SEC11-BP04 Análises manuais de código
- SEC11-BP05 Centralizar serviços para pacotes e dependências
- SEC11-BP06 Implantar software programaticamente
- SEC11-BP07 Avaliar regularmente as propriedades de segurança dos pipelines
- SEC11-BP08 Criar um programa que incorpore a propriedade de segurança nas equipes de workload

SEC11-BP01 Treinar para segurança de aplicações

Forneça treinamento aos criadores em sua organização sobre práticas comuns para promover a segurança no desenvolvimento e na operação de aplicações. A adoção de práticas de desenvolvimento com foco na segurança ajuda a diminuir a probabilidade de problemas que são detectados somente no estágio de avaliação da segurança.

Resultado desejado: o software deve ser projetado e criado com a segurança em mente. Quando os criadores em uma organização são treinados em práticas de desenvolvimento seguras que começam com um modelo de ameaças, isso melhora a qualidade e a segurança gerais do software produzido. Essa abordagem pode reduzir o tempo de entrega do software ou de recursos porque não é necessário tanto retrabalho após o estágio de avaliação da segurança.

Para as finalidades desta prática recomendada, desenvolvimento seguro refere-se ao software que está sendo criado e às ferramentas ou aos sistemas compatíveis com o ciclo de vida de desenvolvimento de software (SDLC).

Antipadrões comuns:

- Aguardar uma avaliação da segurança e, depois, considerar as propriedades de segurança de um sistema.
- Deixar todas as decisões de segurança para a equipe de segurança.
- Não comunicar como as decisões tomadas no SDLC se relacionam às expectativas ou as políticas de segurança gerais da organização.
- Iniciar o processo de avaliação da segurança muito tardiamente.

Benefícios do estabelecimento desta prática recomendada:

- Melhor conhecimento dos requisitos organizacionais para a segurança na fase inicial do ciclo de desenvolvimento.
- Ser capaz de identificar e solucionar possíveis problemas de segurança com maior rapidez, promovendo uma entrega de recursos mais rápida.
- Maior qualidade do software e dos sistemas.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

Ofereça treinamento aos criadores em sua organização. Iniciar um curso sobre modelagem de ameaças é uma boa base para ajudar a treinar para segurança. Preferencialmente, os criadores devem ser capazes de acessar de forma independente as informações relevantes às respectivas workloads. Esse acesso os ajuda a tomar decisões embasadas sobre as propriedades de segurança dos sistemas criados por eles sem a necessidade de solicitar outra equipe. O processo para envolver a equipe de segurança para avaliações deve ser claramente definido e simples de seguir. As etapas do processo de avaliação devem ser incluídas no treinamento de segurança. Quando houver padrões ou modelos de implementação disponíveis, eles deverão ser simples de encontrar e vincular aos requisitos de segurança gerais. Considere usar o AWS CloudFormation, as estruturas do AWS Cloud Development Kit (AWS CDK), o Service Catalog ou outras ferramentas de modelo para reduzir a necessidade de configuração personalizada.

Etapas da implementação

- Oferecer aos criadores um curso sobre modelagem de ameaças para criar uma boa base e ajudar a treiná-los a pensar em segurança.
- Conceder acesso ao treinamento do Treinamento da AWS and Certification, do setor ou de parceiros da AWS.
- Fornecer treinamento sobre o processo de avaliação da segurança de sua organização, que esclarece a divisão de responsabilidades entre a equipe de segurança, as equipes de workload e outras partes interessadas.
- Publicar orientações de autoatendimento sobre como atender aos seus requisitos de segurança, inclusive códigos de exemplo e modelos, se disponíveis.
- Obter feedback regularmente de equipes de criadores sobre a experiência deles com o processo e o treinamento de processo de avaliação da segurança e usar esse feedback para promover melhorias.
- Utilizar dias de jogo ou campanhas de bug bash para ajudar a reduzir o número de problemas e aumentar as habilidades de seus criadores.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP08 Criar um programa que incorpore a propriedade de segurança nas equipes de workload

Documentos relacionados:

- Treinamento da AWS and Certification
- Como pensar sobre governança de segurança na nuvem
- Como abordar a modelagem de ameaças
- Como acelerar o treinamento: o AWS Skills Guild

Vídeos relacionados:

- Segurança proativa: considerações e abordagens

Exemplos relacionados:

- Workshop sobre modelagem de ameaças
- Conscientização do setor para desenvolvedores

Serviços relacionados:

- AWS CloudFormation
- Estruturas do AWS Cloud Development Kit (AWS CDK) (AWS CDK)
- Service Catalog
- AWS BugBust

SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Automatize o teste das propriedades de segurança durante o ciclo de vida de desenvolvimento e lançamento. Com a automação, é mais fácil identificar de forma consistente e repetível possíveis problemas no software antes do lançamento, o que reduz o risco de problemas de segurança no software que está sendo fornecido.

Resultado esperado: o objetivo do teste automatizado é oferecer uma forma programática de detectar possíveis problemas precocemente e com frequência ao longo do ciclo de vida de desenvolvimento. Ao automatizar o teste de regressão, você pode executar novamente testes funcionais e não funcionais para verificar se o software testado anteriormente ainda funciona da forma esperada após uma alteração. Ao definir testes de unidade de segurança para conferir

configurações incorretas comuns, como uma autenticação ausente ou danificada, é possível identificar e resolver esses problemas logo no início do processo de desenvolvimento.

A automação de testes utiliza casos de teste para um propósito específico para validação de aplicações, com base nos requisitos e na funcionalidade desejada da aplicação. O resultado dos testes automatizados baseia-se na comparação da saída do teste gerado com a respectiva saída esperada, o que acelera o ciclo de vida dos testes em geral. As metodologias de teste, como teste de regressão e pacotes de teste de unidade, são mais adequadas para automação. A automação dos testes de propriedades de segurança possibilita aos criadores receber feedback automatizado sem precisar esperar por uma avaliação da segurança. Os testes automatizados em forma de análise de código estático ou dinâmico podem melhorar a qualidade do código e ajudar a detectar possíveis problemas de software no ciclo de vida de desenvolvimento.

Antipadrões comuns:

- Não comunicar os casos de teste e os resultados dos testes automatizados.
- Realizar os testes automatizados somente antes de um lançamento.
- Automatizar casos de teste com requisitos que mudam com frequência.
- Não fornecer orientações sobre como abordar os resultados dos testes de segurança.

Benefícios do estabelecimento desta prática recomendada:

- Redução da dependência de pessoas que avaliam as propriedades de segurança dos sistemas.
- Descobertas consistentes em vários fluxos de trabalho que melhoram a consistência.
- Redução da probabilidade de introduzir problemas de segurança no software de produção.
- Redução do período de tempo entre a detecção e a correção devido à detecção mais antecipada de problemas de software.
- Maior visibilidade do problema sistêmico ou repetido entre os vários fluxos de trabalho, o que pode ser utilizado para promover melhorias em toda a organização.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientação de implementação

Ao criar um software, adote vários mecanismos de teste para garantir que você esteja testando os requisitos funcionais da aplicação, com base na respectiva lógica de negócios e em requisitos não funcionais, os quais se concentram na confiabilidade, performance e segurança da aplicação.

O teste de segurança de aplicação estática (SAST) analisa padrões de segurança anômalos no código-fonte e fornece indicações de código propenso a defeitos. O SAST depende de entradas estáticas, como documentação (especificação de requisitos, documentação e especificações de design) e código-fonte da aplicação, para testar uma série de problemas de segurança conhecidos. Os analisadores de código estático podem ajudar a acelerar a análise de grandes volumes de código. O NIST Quality Group oferece uma comparação de analisadores de segurança de código-fonte, o que inclui ferramentas de código aberto para leitores de código de byte e leitores de código binário.

Complemente seu teste estático com metodologias de teste de segurança de análise dinâmica (DAST), que realizam testes na aplicação em execução a fim de identificar comportamento possivelmente inesperado. O teste dinâmico pode ser utilizado para detectar possíveis problemas que não são detectáveis por meio de análise estática. Por meio dos testes nos estágios de repositório de código, compilação e pipeline, é possível impedir que diferentes tipos de problema em potencial ocorram no código. O Amazon CodeWhisperer oferece recomendações de código, como verificação de segurança, no IDE do criador. O Amazon CodeGuru Reviewer pode identificar problemas críticos, problemas de segurança e bugs difíceis de detectar durante o desenvolvimento da aplicação e oferece recomendações para melhorar a qualidade do código.

O workshop Segurança para desenvolvedores utiliza ferramentas de desenvolvedor da AWS, como AWS CodeBuild, AWS CodeCommit e AWS CodePipeline, para automação de pipeline de lançamento que inclui as metodologias de teste SAST e DAST.

À medida que você avançar no SDLC, estabeleça um processo iterativo que inclua avaliações de aplicação periódicas com sua equipe de segurança. O feedback coletado dessas avaliações de segurança deve ser abordado e validado como parte de sua avaliação de prontidão do lançamento. Essas avaliações estabelecem um procedimento de segurança robusto de aplicações e fornecem aos criadores feedback útil para resolver possíveis problemas.

Etapas da implementação

- Implementar um IDE consistente, análise de código e ferramentas de CI/CD que incluam teste de segurança.

- Considerar quando no SDLC é adequado bloquear pipelines em vez de apenas notificar os criadores de que problemas precisam ser corrigidos.
- O workshop Segurança para desenvolvedores fornece um exemplo de como integrar testes estáticos e dinâmicos a um pipeline de lançamento.
- Realizar testes ou análise de código com ferramentas automatizadas, como o Amazon CodeWhisperer integrado a IDEs de desenvolvedores e o Amazon CodeGuru Reviewer para verificação do código na confirmação, ajuda os criadores a obter feedback no momento certo.
- Ao criar com o AWS Lambda, é possível usar o Amazon Inspector para verificar o código de aplicação em suas funções.
- O workshop CI/CD na AWS fornece um ponto de partida para criar pipelines de CI/CD na AWS.
- Quando testes automatizados são incluídos em pipelines de CI/CD, você precisa usar um sistema de emissão de tíquetes para rastrear a notificação e a correção de problemas de software.
- Para testes de segurança que podem gerar descobertas, a vinculação com orientações para correção ajuda os criadores a melhorar a qualidade do código.
- Analise regularmente as descobertas das ferramentas automatizadas para priorizar a próxima automação, o treinamento de criadores ou a campanha de conscientização.

Recursos

Documentos relacionados:

- Entrega contínua e implantação contínua

- Parceiros com competência em DevOps da AWS
- Parceiros de competência em segurança da AWS para segurança da aplicação
- Como escolher uma abordagem de CI/CD do Well-Arquitected
- Monitorar eventos do CodeCommit no Amazon EventBridge e no Amazon CloudWatch Events
- Análise da detecção de segredos no Amazon CodeGuru Review
- Acelerar implantações na AWS com governança efetiva
- Como a AWS aborda a automação de implantações seguras e sem intervenção manual

Vídeos relacionados:

- Sem intervenção manual: como automatizar os pipelines de entrega contínua na Amazon

- Como automatizar pipelines CI/CD entre contas

Exemplos relacionados:

- Conscientização do setor para desenvolvedores
- Governança do AWS CodePipeline
- Workshop Segurança para desenvolvedores
- Workshop sobre CI/CD da AWS

SEC11-BP03 Realizar teste de penetração regular

Realize teste de penetração regular do software. Esse mecanismo ajuda a identificar possíveis problemas de software que não podem ser detectados pelo teste automatizado ou por uma análise manual do código. Ele também ajuda você a entender a eficácia dos controles de detecção. O teste de penetração deve tentar determinar se o software pode ser executado de formas inesperadas; por exemplo, expondo dados que devem ser protegidos ou concedendo permissões mais amplas que o esperado.

Resultado desejado: o teste de penetração é usado para detectar, corrigir e validar as propriedades de segurança da aplicação. O teste de penetração regular e programado deve ser realizado como parte do ciclo de vida de desenvolvimento de software (SDLC). As descobertas do teste de penetração devem ser abordadas antes do lançamento do software. Você precisa analisar as descobertas do teste de penetração para identificar se há problemas que podem ser encontrados usando a automação. Ter um processo de teste de penetração regular e repetível que inclua um mecanismo de feedback ativo ajuda a transmitir as orientações aos criadores e melhora a qualidade do software.

Antipadrões comuns:

- Realizar um teste de penetração somente para problemas de segurança conhecidos ou prevalentes.
- Realizar um teste de penetração em aplicações sem ferramentas e bibliotecas de terceiro dependentes.
- Realizar um teste de penetração em aplicações em busca de problemas de segurança de pacote e não avaliar a lógica de negócios implementada.

Benefícios do estabelecimento desta prática recomendada:

- Maior confiança nas propriedades de segurança do software antes do lançamento.
- Oportunidade de identificar padrões de aplicação preferenciais, o que aumenta a qualidade do software.
- Um ciclo de feedback que identifica mais cedo no ciclo de desenvolvimento quando a automação ou treinamento adicional pode melhorar as propriedades de segurança do software.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: alto

Orientação de implementação

O teste de penetração é um exercício de teste de segurança estruturado em que você executa cenários de violação de segurança planejados a fim de detectar, corrigir e validar controles de segurança. Os testes de penetração começam com o reconhecimento, durante o qual os dados são coletados com base no design atual da aplicação e nas respectivas dependências. Uma lista selecionada de cenários de teste específicos de segurança é criada e executada. A principal finalidade desses testes é revelar problemas de segurança em sua aplicação, que podem ser explorados para obter acesso não intencional ao seu ambiente ou acesso não autorizado aos dados. Você precisa realizar o teste de penetração ao lançar novos recursos ou sempre que sua aplicação passar por alterações importantes na implementação técnica ou de funções.

É necessário identificar o estágio mais apropriado do ciclo de vida de desenvolvimento para realizar o teste de penetração. Esse teste deve ocorrer em uma fase tardia o suficiente para que a funcionalidade do sistema esteja próxima ao estado de lançamento pretendido, mas com tempo suficiente para corrigir todos os problemas.

Etapas da implementação

- Ter um processo estruturado sobre como definir o escopo do teste de penetração. Basear esse processo no modelo de ameaças é uma boa forma de manter o contexto.
- Identificar o estágio apropriado do ciclo de vida de desenvolvimento para realizar o teste de penetração, que deve ser quando houver o mínimo de alterações esperadas na aplicação e houver tempo suficiente para realizar a correção.
- Treinar os criadores sobre o que esperar das descobertas do teste de penetração e como ter informações sobre correção.
- Utilizar ferramentas para acelerar o processo de testes de penetração automatizando testes comuns ou repetíveis.

- Analisar as descobertas do teste de penetração para identificar problemas de segurança sistêmicos e utilizar esses dados para embasar testes automatizados adicionais e a instrução contínua dos criadores.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP01 Treinar para segurança de aplicações
- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Documentos relacionados:

- O teste de penetração da AWS fornece orientações detalhadas para teste de penetração na AWS
- Acelerar implantações na AWS com governança efetiva
- Parceiros de competência em segurança da AWS
- Modernize sua arquitetura de teste de penetração no AWS Fargate
- AWS Fault injection Simulator

Exemplos relacionados:

- Como automatizar testes de API com o AWS CodePipeline (GitHub)
- Assistente de segurança automatizado (GitHub)

SEC11-BP04 Análises manuais de código

Realize uma análise manual do código do software que você produz. Esse processo ajuda a verificar se a pessoa que escreveu o código não é a única que está conferindo a qualidade dele.

Resultado desejado: a inclusão de uma etapa de análise de código manual durante o desenvolvimento melhora a qualidade do software que está sendo criado, ajuda a melhorar as habilidades de membros menos experientes da equipe e oferece uma oportunidade de identificar locais onde a automação pode ser usada. É possível oferecer compatibilidade com as análises de código manuais com ferramentas e testes automatizados.

Antipadrões comuns:

- Não realizar análises de código antes da implantação.
- Ter a mesma pessoa para escrever e analisar o código.
- Não utilizar a automação para auxiliar ou orquestrar as análises de código.
- Não treinar os criadores em segurança de aplicações antes de analisarem o código.

Benefícios do estabelecimento desta prática recomendada:

- Código de melhor qualidade.
- Maior consistência do desenvolvimento do código por meio da reutilização de abordagens comuns.
- Redução no número de problemas descobertos durante o teste de penetração e em estágios posteriores.
- Maior transferência de conhecimentos na equipe.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

A etapa de análise deve ser implementada como parte do fluxo de gerenciamento de código geral. Os detalhes dependem da abordagem utilizada para ramificação, solicitações de pull e mesclagem. Você pode utilizar o AWS CodeCommit ou soluções de terceiros, como GitHub, GitLab ou Bitbucket. Seja qual for o método utilizado, é importante verificar se seus processos precisam de análise de código antes da implantação em um ambiente de produção. O uso de ferramentas, como o Amazon CodeGuru Reviewer, pode facilitar a orquestração do processo de análise do código.

Etapas da implementação

- Implementar uma etapa de análise manual como parte do fluxo de gerenciamento de código e realizar essa análise antes de prosseguir.
- Considerar o Amazon CodeGuru Reviewer para gerenciar e auxiliar nas análises de código.
- Implementar um fluxo de aprovação que exija a realização de uma análise de código antes de avançá-lo para o próximo estágio.
- Verificar se há um processo para identificar problemas encontrados durante as análises de código manuais que possam ser detectados automaticamente.
- Integrar a etapa de análise de código manual de forma que se alinhe às suas práticas de desenvolvimento de código.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Documentos relacionados:

- Trabalhar com solicitações de pull em repositórios do AWS CodeCommit
- Trabalhar com modelos de regra de aprovação no AWS CodeCommit
- Sobre solicitações de pull no GitHub
- Análises de código automatizadas com o Amazon CodeGuru Reviewer
- Automatizar a detecção de vulnerabilidades de segurança e bugs em pipelines de CI/CD com o uso da CLI do Amazon CodeGuru Reviewer

Vídeos relacionados:

- Melhoria contínua da qualidade do código com o Amazon CodeGuru

Exemplos relacionados:

- Workshop Segurança para desenvolvedores

SEC11-BP05 Centralizar serviços para pacotes e dependências

Forneça serviços centralizados a equipes de criadores para obter pacotes de software e outras dependências. Isso permite a validação de pacotes antes que eles sejam incluídos no software que você escreve e fornece uma fonte de dados para a análise do software que está sendo usado na sua organização.

Resultado desejado: o software é composto de um conjunto de outros pacotes de software além do código que está sendo escrito. Isso simplifica o consumo de implementações de funcionalidades que são utilizadas repetidamente, como um analisador JSON ou uma biblioteca de criptografia. A centralização lógica das fontes desses pacotes e dependências oferece um mecanismo para as equipes de segurança validarem as propriedades dos pacotes antes de eles serem utilizados. Essa abordagem também reduz o risco de um problema inesperado ser provocado por uma alteração em um pacote existente ou pela inclusão de pacotes arbitrários diretamente da Internet pelas equipes

de criadores. Utilize essa abordagem em conjunto com os fluxos de testes manuais e automatizados para aumentar a confiança na qualidade do software que está sendo desenvolvido.

Antipadrões comuns:

- Extrair pacotes de repositórios arbitrários na Internet.
- Não testar novos pacotes antes de disponibilizá-los aos criadores.

Benefícios do estabelecimento desta prática recomendada:

- Melhor entendimento de quais pacotes estão sendo utilizados no software que está sendo criado.
- Capacidade de notificar as equipes de workload quando um pacote precisa ser atualizado com base no entendimento de quem está usando o quê.
- Redução do risco de um pacote com problemas ser incluído em seu software.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientação de implementação

Forneça serviços centralizados para pacotes e dependências de uma forma simples para os criadores consumirem. Serviços centralizados podem ser centralizados logicamente em vez de implementados como um sistema monolítico. Essa abordagem possibilita fornecer serviços de uma forma que atenda às necessidades dos criadores. Você precisa implementar uma forma eficiente de adicionar pacotes ao repositório quando ocorrem atualizações ou surgem novos requisitos. Serviços da AWS como o AWS CodeArtifact ou soluções semelhantes de parceiros da AWS oferecem uma forma de entregar esse recurso.

Etapas da implementação:

- Implementar um serviço de repositório centralizado logicamente disponível em todos os ambientes onde o software é desenvolvido.
- Incluir acesso ao repositório como parte do processo de provisionamento de Conta da AWS.
- Criar automação para testar pacotes antes de serem publicados em um repositório.
- Manter métricas dos pacotes mais utilizados, das linguagens e das equipes com a maior quantidade de alterações.
- Fornecer um mecanismo automatizado para as equipes de criadores solicitarem novos pacotes e fornecerem feedback.

- Verificar regularmente os pacotes em seu repositório para identificar o possível impacto de problemas recém-descobertos.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Documentos relacionados:

- Acelerar implantações na AWS com governança efetiva
- Aumentar a segurança de seu pacote com o kit de ferramentas CodeArtifact Package Origin Control
- Detectar problemas de segurança no registro em log com o Amazon CodeGuru Reviewer
- Níveis de cadeia de suprimentos para artefatos de software (SLSA)

Vídeos relacionados:

- Segurança proativa: considerações e abordagens
- A filosofia de segurança da AWS (re:Invent 2017)
- Quando a segurança, a proteção e a urgência importam: lidar com o Log4Shell

Exemplos relacionados:

- Pipeline de publicação de pacotes de várias regiões (GitHub)
- Publicar módulos Node.Js no AWS CodeArtifact usando o AWS CodePipeline (GitHub)
- Exemplo de pipeline do AWS CDK Java CodeArtifact (GitHub)
- Distribuir pacotes privados do .NET NuGet com o AWS CodeArtifact (GitHub)

SEC11-BP06 Implantar software programaticamente

Faça implantações de software de forma programática quando possível. Essa abordagem diminui a probabilidade de falha em uma implantação ou da introdução de um problema inesperado devido a erro humano.

Resultado desejado: manter as pessoas longe dos dados é um princípio essencial da criação segura na Nuvem AWS. Esse princípio inclui como implantar seu software.

Os benefícios de não contar com pessoas para implantar software é a maior confiança de que o componente testado é o que será implantado e de que a implantação sempre é realizada de forma consistente. O software não deve precisar de alterações para funcionar em diferentes ambientes. O uso dos princípios de desenvolvimento de aplicações de 12 fatores, especificamente a externalização da configuração, possibilita implantar o mesmo código em vários ambientes sem a necessidade de alterações. Assinar de forma criptográfica os pacotes de software é uma boa maneira de garantir que nada tenha sido alterado entre os ambientes. O resultado geral dessa abordagem é reduzir o risco em seu processo de alterações e melhorar a consistência das versões do software.

Antipadrões comuns:

- Implantar software manualmente em produção.
- Realizar alterações manualmente no software para suprir diferentes ambientes.

Benefícios do estabelecimento desta prática recomendada:

- Maior confiança no processo de lançamento de software.
- Redução do risco de uma alteração com falha afetar a funcionalidade dos negócios.
- Maior cadência de lançamentos devido ao menor risco de alterações.
- Recurso de reversão automática para eventos inesperados durante a implantação.
- Capacidade de comprovar de forma criptográfica que o software testado é o software implantado.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: alto

Orientações para a implementação

Crie a infraestrutura de sua Conta da AWS para remover o acesso humano persistente dos ambientes e use ferramentas de CI/CD para realizar implantações. Projete suas aplicações de forma que os dados da configuração específica do ambiente sejam obtidos de uma fonte externa, como o AWS Systems Manager Parameter Store. Assine pacotes depois de testados e valide essas assinaturas durante a implantação. Configure seus pipelines de CI/CD para enviar código da aplicação e usar canários para confirmar a implantação bem-sucedida. Utilize ferramentas como o

AWS CloudFormation ou o AWS CDK para definir sua infraestrutura; depois, use o AWS CodeBuild e o AWS CodePipeline para realizar operações de CI/CD.

Etapas da implementação

- Criar pipelines de CI/CD bem definidos para simplificar o processo de implantação.
- O uso do AWS CodeBuild e do AWS Code Pipeline para oferecer recurso de CI/CD simplifica a integração de teste de segurança aos seus pipelines.
- Seguir as orientações sobre separação de ambientes no whitepaper Organizar seu ambiente da AWS com o uso de várias contas.
- Garantir que não haja nenhum acesso humano persistente aos ambientes nos quais as workloads de produção estão em execução.
- Projetar as aplicações para oferecer compatibilidade com a externalização de dados de configuração.
- Considerar a implantação com o uso do modelo de implantação azul/verde.
- Implementar canários para validar a implantação bem-sucedida do software.
- Utilizar ferramentas criptográficas, como o AWS Signer ou o AWS Key Management Service (AWS KMS), para assinar e confirmar os pacotes de software que você está implantando.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Documentos relacionados:

- Workshop sobre CI/CD da AWS
- Acelerar implantações na AWS com governança efetiva
- Automatizar uma implantação prática e sem intervenção manual
- Assinatura de código com o uso de CA privada do AWS Certificate Manager e chaves assimétricas do AWS Key Management Service)
- Assinatura de código: um controle de integridade e confiança para o AWS Lambda

Vídeos relacionados:

- Sem intervenção manual: como automatizar os pipelines de entrega contínua na Amazon

Exemplos relacionados:

- Implantações azul/verde com o AWS Fargate

SEC11-BP07 Avaliar regularmente as propriedades de segurança dos pipelines

Aplique os princípios do pilar Segurança do Well-Architected aos seus pipelines, com atenção especial à separação das permissões. Avalie as propriedades de segurança de sua infraestrutura de pipelines. O gerenciamento eficaz da segurança dos pipelines permite que você forneça segurança ao software que passa pelos pipelines.

Resultado desejado: os pipelines utilizados para criar e implantar o software devem seguir as mesmas práticas recomendadas que qualquer outra workload em seu ambiente. Os testes implementados nos pipelines não devem ser editáveis pelos criadores que os estão utilizando. Os pipelines só devem ter as permissões necessárias para as implantações que eles estão realizando e devem implementar proteções para evitar a implantação em ambientes errados. Os pipelines não devem contar com credenciais de longo prazo e devem ser configurados para emitir o estado de forma que a integridade dos ambientes de compilação possa ser validada.

Antipadrões comuns:

- Testes de segurança que podem ser ignorados pelos criadores.
- Permissões excessivamente amplas para pipelines de implantação.
- Pipelines não configurados para validar entradas.
- Ausência de análise regular das permissões associadas à infraestrutura de CI/CD.
- Uso de credenciais de longo prazo ou codificadas.

Benefícios do estabelecimento desta prática recomendada:

- Maior confiança na integridade do software que está sendo criado e implantado pelos pipelines.
- Capacidade de interromper uma implantação quando há atividade suspeita.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: alto

Orientação de implementação

Iniciar com serviços de CI/CD gerenciados que ofereçam compatibilidade com perfis do IAM reduz o risco de vazamento de credenciais. Aplicar os princípios do pilar Segurança à sua infraestrutura de pipeline de CI/CD pode ajudar você a determinar onde é possível realizar melhorias de segurança. Seguir a Arquitetura de referência de pipelines de implantação da AWS é um bom ponto de partida para criar seus ambientes de CI/CD. Analisar regularmente a implementação de pipelines e analisar comportamentos inesperados nos logs pode ajudar você a entender os padrões de uso dos pipelines que estão sendo utilizados para implantar o software.

Etapas da implementação

- Iniciar com a Arquitetura de referência de pipeline de implantação da AWS.
- Considerar o uso do AWS IAM Access Analyzer para gerar de forma programática as políticas de privilégio mínimo do IAM para os pipelines.
- Integrar seus pipelines ao monitoramento e aos alertas de forma que você seja notificado de atividade inesperada ou anormal. Para serviços gerenciados da AWS, o Amazon EventBridge possibilita rotear dados para destinos, como o AWS Lambda ou o Amazon Simple Notification Service (Amazon SNS).

Recursos

Documentos relacionados:

- Arquitetura de referência de pipeline de implantação da AWS
- Monitorar o AWS CodePipeline
- Práticas recomendadas de segurança para o AWS CodePipeline

Exemplos relacionados:

- Painel de monitoramento de DevOps (GitHub)

SEC11-BP08 Criar um programa que incorpore a propriedade de segurança nas equipes de workload

Crie um programa ou mecanismo que capacite as equipes de criadores a tomar decisões de segurança sobre o software que elas estão criando. Ainda assim é necessário que sua equipe de

segurança valide essas decisões durante uma avaliação, mas a incorporação da propriedade de segurança nas equipes de criadores aumenta a velocidade e segurança do processo de criação de workloads. Esse mecanismo também promove uma cultura de propriedade que afeta de forma positiva a operação dos sistemas que você cria.

Resultado desejado: para incorporar a propriedade de segurança e a tomada de decisão às equipes de criadores, você pode treinar os criadores a pensar sobre segurança ou incrementar o treinamento deles com pessoal de segurança incorporado ou associado às equipes de criadores. As duas abordagens são válidas e possibilitam à equipe tomar decisões de segurança de melhor qualidade logo no início do ciclo de desenvolvimento. Esse modelo de propriedade é baseado em treinamento para segurança de aplicações. Iniciar com o modelo de ameaças para a workload específica ajuda a direcionar o design thinking (pensamento de design) para o contexto apropriado. Outro benefício de ter uma comunidade de criadores concentrados em segurança ou um grupo de engenheiros de segurança que trabalhem com equipes de criadores é que você pode entender mais profundamente como o software é escrito. Esse entendimento ajuda você a determinar as próximas áreas de melhoria em seu recurso de automação.

Antipadrões comuns:

- Deixar todas as decisões de design de segurança para a equipe de segurança.
- Não abordar os requisitos de segurança cedo o suficiente no processo de desenvolvimento.
- Não obter feedback dos criadores e do pessoal de segurança sobre a operação do programa.

Benefícios do estabelecimento desta prática recomendada:

- Redução do tempo para concluir as avaliações de segurança.
- Redução dos problemas de segurança que são detectados apenas no estágio de avaliação da segurança.
- Melhoria da qualidade geral do software que está sendo escrito.
- Oportunidade de identificar e entender problemas sistêmicos ou áreas de melhoria de alto valor.
- Redução da quantidade de revisão necessária devido às descobertas da avaliação da segurança.
- Melhoria da percepção da função de segurança.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: baixo

Orientações para a implementação

Comece com as orientações em SEC11-BP01 Treinar para segurança de aplicações. Depois, identifique o modelo operacional para o programa que você acredita ser o melhor para a sua organização. Os dois padrões principais são treinar os criadores ou incorporar o pessoal de segurança às equipes de criadores. Depois de decidir sobre a abordagem inicial, você precisa criar um piloto com uma equipe de workload ou um grupo pequeno de equipes de workload para comprovar que o modelo funciona para sua organização. O apoio de liderança dos criadores e da segurança da organização contribui para a entrega e o sucesso do programa. À medida que você criar esse programa, é importante selecionar as métricas que podem ser utilizadas para mostrar o valor dele. Saber como a AWS resolveu esse problema é uma boa experiência de aprendizado. A prática recomendada é muito concentrada na mudança e cultura organizacionais. As ferramentas que você utiliza devem ser compatíveis com a colaboração entre as comunidades de criadores e de segurança.

Etapas da implementação

- Começar com o treinamento dos criadores para segurança de aplicações.
- Criar uma comunidade e um programa de integração para instruir os criadores.
- Selecionar um nome para o programa. Guardiões, patrocinadores ou defensores são utilizados com frequência.
- Identificar o modelo a ser utilizado: treinar criadores, incorporar engenheiros de segurança e ter perfis de segurança de afinidade.
- Identificar patrocinadores do projeto em grupos de segurança e de criadores e possivelmente em outros grupos relevantes.
- Rastrear as métricas do número de pessoas envolvidas no programa, o tempo gasto em avaliações e o feedback dos criadores e do pessoal de segurança. Utilizar essas métricas para realizar melhorias.

Recursos

Práticas recomendadas relacionadas:

- SEC11-BP01 Treinar para segurança de aplicações
- SEC11-BP02 Automatizar o teste durante o ciclo de vida de desenvolvimento e lançamento

Documentos relacionados:

- Como abordar a modelagem de ameaças
- Como pensar sobre governança de segurança na nuvem

Vídeos relacionados:

- Segurança proativa: considerações e abordagens

Confiabilidade

O pilar Confiabilidade abrange a capacidade de uma workload de executar a função pretendida correta e consistentemente quando esperado. Você pode encontrar orientações prescritivas sobre implementação no whitepaper sobre o pilar de confiabilidade.

Áreas de práticas recomendadas

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

Fundamentos

Perguntas

- CONFIABILIDADE 1. Como gerenciar as Service Quotas e restrições?
- CONFIABILIDADE 2. Como planejar sua topologia de rede?

CONFIABILIDADE 1. Como gerenciar as Service Quotas e restrições?

Para arquiteturas de workload baseadas na nuvem, há Service Quotas, que também são conhecidas como limites de serviço. Essas cotas existem para evitar o provisionamento 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 restrições e as cotas fixas de serviço por meio da arquitetura
- REL01-BP04 Monitorar e gerenciar cotas
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL01-BP06 Garantir 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

Esteja ciente das suas cotas padrão e das solicitações de aumento de cota referentes à sua arquitetura de workload. Saiba quais restrições de recursos, como disco ou rede, podem gerar impactos.

Resultado desejado: os clientes conseguem evitar a degradação ou a interrupção do serviço nas Contas da AWS implementando diretrizes adequadas para monitorar as principais métricas, análises da infraestrutura e etapas de remediação da automação, a fim de confirmar que as cotas e as restrições do serviço não foram atingidas, o que poderia causar degradação ou interrupção do serviço.

Antipadrões comuns:

- Implantar uma workload sem compreender as cotas flexíveis ou fixas e seus limites para os serviços utilizados.
- Implantar uma workload de substituição sem analisar e reconfigurar as cotas necessárias ou entrar em contato com o suporte com antecedência.
- Pressupor que os serviços em nuvem não têm limites e os serviços podem ser usados sem considerar taxas, limites, contagens e quantidades.
- Pressupor que as cotas aumentarão automaticamente.
- Não saber o processo e a linha de tempo das solicitações de cota.
- Pressupor que a cota de serviço em nuvem padrão é idêntica para todos os serviços em comparação entre as regiões.
- Pressupor que as restrições do serviço podem ser violadas e os sistemas vão ser escalados automaticamente ou aumentar o limite além das restrições do recurso
- Não testar a aplicação em tráfego de pico a fim de aplicar tensão na utilização de seus recursos.

- Provisionar o recurso sem analisar o tamanho necessário dele.
- Superprovisionar capacidade selecionando tipos de recurso que superam em muito a necessidade real ou os picos esperados.
- Não avaliar os requisitos de capacidade para novos níveis de tráfego antes de um novo evento de cliente ou implantação de uma nova tecnologia.

Benefícios do estabelecimento desta prática recomendada: o monitoramento e o gerenciamento automatizado de cotas de serviço e restrições de recursos podem reduzir as falhas de forma proativa. As alterações nos padrões de tráfego do serviço de um cliente poderão causar interrupção ou degradação se as práticas recomendadas não forem seguidas. Ao monitorar e gerenciar esses valores em todas as regiões e contas, as aplicações podem ter uma resiliência aprimorada em eventos adversos ou não planejados.

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

Orientações para a implementação

O Service Quotas é um serviço da AWS que ajuda você a gerenciar as cotas de mais de 250 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 SDK. 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 por respectivo serviço, por exemplo, consulte Cotas da Amazon VPC.

Alguns limites de serviço, como os limites de taxa para APIs limitadas, são definidos no próprio Amazon API Gateway por meio da configuração de um plano de uso. Alguns limites definidos como configuração em seus respectivos serviços incluem IOPS provisionadas, armazenamento do Amazon RDS alocado e alocações de volume do Amazon EBS. O Amazon Elastic Compute Cloud tem seu próprio painel de limites de serviço, que pode ajudar você a gerenciar sua instância, o Amazon Elastic Block Store e os limites de endereços IP elásticos. Se você tiver um caso de uso em que as cotas de serviço afetam a performance de sua aplicação e elas não forem ajustadas às suas necessidades, entre em contato com o AWS Support para ver se há mitigações.

As cotas de serviço podem ser específicas da região e também pode ser globais por natureza. O uso de um serviço da AWS com a cota atingida fará com que o comportamento dele não seja o esperado e poderá causar interrupção ou degradação do serviço. Por exemplo, a cota de um serviço limita o número de DL Amazon EC2 que pode ser usado em uma região e esse limite poderá ser atingido durante um evento de escalabilidade de tráfego usando grupos do Auto Scaling (ASG).

As cotas de serviço de cada conta devem ser avaliadas regularmente quanto ao uso a fim de determinar quais são os limites de serviço apropriados para a conta em questão. Essas cotas de serviço existem como barreiras de proteção operacionais, a fim de impedir o provisionamento acidental de recursos além do necessário. Elas também servem para limitar as taxas de solicitação em operações de API para proteger os serviços contra abuso.

Restrições de serviço são diferentes de cotas de serviço. As restrições de serviço representam os limites de um recurso específico conforme definido pelo tipo de recurso em questão. Podem ser a capacidade de armazenamento (por exemplo, o gp2 tem um limite de tamanho de 1 GB a 16 TB) ou o throughput de disco (10 mil iops). É essencial que a restrição de um tipo de recurso seja projetada e avaliada constantemente quanto ao uso que pode atingir o limite. Se uma restrição for atingida de modo inesperado, as aplicações da conta ou os serviços poderão sofrer degradação ou interrupção.

Se houver um caso de uso em que as cotas de serviço afetem a performance de uma aplicação e elas não puderem ser ajustadas às necessidades, entre em contato com o AWS Support para ver se há mitigações. Para obter mais detalhes sobre o ajuste de cotas fixas, consulte REL01-BP03 Acomodar as restrições e as cotas fixas de serviço por meio da arquitetura.

Há uma série de serviços e ferramentas da AWS para ajudar a monitorar e gerenciar o Service Quotas. O serviço e as ferramentas devem ser utilizadas para oferecer verificações automatizadas ou manuais dos níveis de cota.

- O AWS Trusted Advisor oferece uma verificação de cotas de serviço que exibe o uso e cotas para alguns aspectos de alguns serviços. Ele pode ajudar na identificação de serviços que estão próximos da cota.
- O AWS Management Console oferece métodos para exibir valores de cota de serviço, gerenciar, solicitar novas cotas, monitorar o status das solicitações de cota e exibir o histórico de cotas.
- A AWS CLI e os CDKs oferecem métodos programáticos para gerenciar e monitorar automaticamente os níveis e o uso de cotas de serviço.

Etapas da implementação

Para Service Quotas:

- Analisar o AWS Service Quotas.
- Para saber suas cotas de serviço existentes, determine os serviços (como o IAM Access Analyzer) utilizados. Há cerca de 250 serviços da AWS controlados por cotas de serviço. Depois, determine

o nome específico da cota de serviço que pode estar sendo usada em cada conta e região. Há cerca de 3 mil nomes de cota de serviço por região.

- Incremente essa análise de cota com o AWS Config para encontrar todos os recursos da AWS utilizados em suas Contas da AWS.
- Use dados do AWS CloudFormation para determinar seus recursos da AWS utilizados. Examine os recursos que foram criados no AWS Management Console ou com o comando `list-stack-resources` da AWS CLI. Você também pode ver no próprio modelo os recursos configurados para implantação.
- Examine o código da implantação para determinar todos os serviços necessários à sua workload.
- Determine as cotas de serviço aplicáveis. Use as informações acessíveis programaticamente por meio do Trusted Advisor e do Service Quotas.
- Estabeleça um método de monitoramento automatizado (consulte REL01-BP02 Gerenciar cotas de serviço de várias contas e regiões e REL01-BP04 Monitorar e gerenciar cotas) para alertar e informar se as cotas de serviço estiverem perto do limite ou o atingirem.
- Estabeleça um método automatizado e programático para conferir se uma cota de serviço foi alterada em uma região, mas não em outras na mesma conta (consulte REL01-BP02 Gerenciar cotas de serviço de várias contas e regiões e REL01-BP04 Monitorar e gerenciar cotas).
- Automatize as verificações de logs e métricas de aplicações para determinar se há erros de restrição de serviço ou cota. Se houver esses erros, envie alertas ao sistema de monitoramento.
- Estabeleça os procedimentos de engenharia para calcular a alteração necessária na cota (consulte REL01-BP05 Automatizar o gerenciamento de cotas) depois de identificar que são necessárias cotas maiores para serviços específicos.
- Crie um fluxo de trabalho de provisionamento e aprovação para solicitar alterações na cota de serviço. Isso deve incluir um fluxo de trabalho de exceção em caso de negação de solicitação ou aprovação parcial.
- Crie um método de engenharia para analisar cotas de serviço antes de provisionar e usar novos serviços da AWS antes de distribuir na produção ou carregar ambientes (por exemplo, conta de teste de carga).

Para restrições de serviço:

- Estabeleça métodos de monitoramento e métricas para alertar sobre recursos que estejam próximos de suas restrições de recurso. Utilize o CloudWatch conforme apropriado para métricas ou monitoramento de logs.

- Estabeleça limites de alerta para cada recurso que tenha uma restrição significativa para a aplicação ou o sistema.
- Crie um fluxo de trabalho e procedimentos de gerenciamento de infraestrutura para alterar o tipo de recurso se a restrição estiver próxima da utilização. Esse fluxo de trabalho deve incluir testes de carga como prática recomendada para verificar se o novo tipo de recurso é o correto com as novas restrições.
- Migre o recurso identificado para o novo tipo de recurso usando os procedimentos e os processos existentes.

Recursos

Práticas recomendadas relacionadas:

- REL01-BP02 Gerenciar cotas de serviço de várias contas e regiões
- REL01-BP03 Acomodar as restrições e as cotas fixas de serviço por meio da arquitetura
- REL01-BP04 Monitorar e gerenciar cotas
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL01-BP06 Garantir que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover
- REL03-BP01 Escolher como segmentar a workload
- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência por meio da engenharia do caos

Documentos relacionados:

- AWS Pilar Confiabilidade da Well-Architected Framework: Disponibilidade
- AWS Service Quotas (anteriormente chamado de limites de serviço)
- AWS Trusted Advisor Best Practice Checks (Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Service Limits (Limites de serviço))
- AWS limit monitor on AWS answers (Monitor de limites da AWS em respostas da AWS)
- Amazon EC2 Service Limits (Limites de serviço do Amazon EC2)
- What is Service Quotas? (O que é o Service Quotas?)

- How to Request Quota Increase (Como solicitar aumento de cota)
- Service endpoints and quotas (Endpoints e cotas de serviço)
- Guia do usuário do Service Quotas
- Quota Monitor for AWS (Monitor de cotas da AWS)
- AWS Fault Isolation Boundaries (Limites de isolamento de falhas da AWS)
- Availability with redundancy (Disponibilidade com redundância)
- AWS para dados
- O que significa integração contínua?
- O que significa distribuição contínua?
- Parceiro do APN: parceiros que podem ajudar no gerenciamento de configuração
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS (Gerenciar o ciclo devida da conta em ambientes de SaaS de conta por locatário na AWS)
- Managing and monitoring API throttling in your workloads (Gerenciar e monitorar o controle de utilização de API em workloads)
- View AWS Trusted Advisor recommendations at scale with AWS Organizations (Exibir recomendações do AWS Trusted Advisor em grande escala com AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower (Automatizar aumentos de limite de serviço e suporte empresarial com AWS Control Tower)

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Exibir e gerenciar cotas para serviços da AWS usando o Service Quotas)
- AWS IAM Quotas Demo (Demonstração de cotas do AWS IAM)

Ferramentas relacionadas:

- Amazon CodeGuru Reviewer
- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge

- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP02 Gerenciar cotas de serviço de várias contas e regiões

Se você estiver usando várias contas ou regiões, solicite as cotas adequadas em todos os ambientes nos quais suas workloads de produção são executadas.

Resultado desejado: os serviços e as aplicações não devem ser afetados pelo esgotamento da cota de serviço para configurações que abrangem contas ou regiões ou que têm designs de resiliência que usam failover de conta, zona ou região.

Antipadrões comuns:

- Permitir que a utilização de recursos em uma região de isolamento aumente sem nenhum mecanismo para manter a capacidade das demais.
- Configurar manualmente todas as cotas nas regiões de isolamento de forma independente.
- Não considerar o efeito das arquiteturas de resiliência (como ativa ou passiva) em necessidades futuras de cota durante a degradação na região que não é a principal.
- Não avaliar as cotas regularmente e fazer alterações necessárias em cada região e conta nas quais a workload é executada.
- Não utilizar modelos de solicitação de cota para solicitar aumentos em várias regiões e contas.
- Não atualizar as cotas de serviço por imaginar incorretamente que aumentar as cotas tem implicações de custo, como solicitações de reserva computacional.

Benefícios do estabelecimento desta prática recomendada: confirmar que você pode lidar com sua carga atual em contas ou regiões secundárias se os serviços regionais ficarem indisponíveis. Isso pode ajudar a reduzir o número de erros ou níveis de degradações que ocorrem durante a perda da região.

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

Orientações para a implementação

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. Manter um alto grau de resiliência exige que as cotas de serviço sejam avaliadas de forma contínua (sejam elas automatizadas ou manuais).

Com mais workloads abrangendo regiões devido à implementação de designs usando as abordagens Ativo/Ativo, Ativo/Passivo: Quente, Ativo/Passivo: Frio e Ativo/Passivo: Luz piloto, é essencial entender todos os níveis de cota de contas e regiões. Padrões de tráfego passados nem sempre são um bom indicador de que a cota de serviço está definida corretamente.

Igualmente importante, o limite do nome da cota de serviço nem sempre é o mesmo para cada região. Em uma região, o valor pode ser cinco e em outra região pode ser dez. O gerenciamento dessas cotas deve abranger todos os mesmos serviços, contas e regiões para fornecer resiliência consistente sob carga.

Reconcilie todas as diferenças de cota de serviço em todas as diferentes regiões (Região ativa ou Região passiva) e crie processos para reconciliar de forma contínua essas diferenças. Os planos de teste de failovers de região passiva raramente são escalados para a capacidade ativa de pico, o que significa que os exercícios de simulações teóricas e dias de teste podem não encontrar diferenças em cotas de serviço entre regiões e também depois manter os limites corretos.

É muito importante rastrear e avaliar o desvio de cotas de serviço, a condição em que os limites de uma cota de serviço específica são alterados em uma região e não em todas. É necessário pensar em alterar a cota em regiões com tráfego ou que possam ter tráfego.

- Selecione as contas e as regiões relevantes conforme seus requisitos de serviço, de latência, regulatórios e de recuperação de desastres.
- 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. Esses valores devem ser comparados em relação a diferenças.

Etapas da implementação

- Analise os valores do Service Quotas que possam ter ultrapassado um nível de risco de uso. O AWS Trusted Advisor oferece alertas para violações de limite de 80% e 90%.

- Analise os valores de cotas de serviço em todas as regiões passivas (em um design ativo/passivo). Verifique se a carga será executada com êxito em regiões secundárias em caso de falha na região principal.
- Automatize a avaliação se ocorreu algum desvio de cota de serviço entre as regiões na mesma conta e aja adequadamente para alterar os limites.
- Se as unidades organizações (UO) do cliente estiverem estruturadas da forma compatível, os modelos de cota de serviço deverão ser atualizados para refletir alterações em todas as cotas que devem ser aplicadas a várias regiões e contas.
  - Crie um modelo e associe regiões à alteração de cota.
  - Analise todos os modelos de cota de serviço existentes para todas as alterações necessárias (região, limites e contas).

Recursos

Práticas recomendadas relacionadas:

- REL01-BP01 Conhecimento das cotas e restrições de serviço
- REL01-BP03 Acomodar as restrições e as cotas fixas de serviço por meio da arquitetura
- REL01-BP04 Monitorar e gerenciar cotas
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL01-BP06 Garantir que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover
- REL03-BP01 Escolher como segmentar a workload
- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência por meio da engenharia do caos

Documentos relacionados:

- AWS Pilar Confiabilidade da Well-Architected Framework: Disponibilidade
- AWS Service Quotas (anteriormente chamado de limites de serviço)
- AWS Trusted Advisor Best Practice Checks (Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Service Limits (Limites de serviço))

- AWS limit monitor on AWS answers (Monitor de limites da AWS em respostas da AWS)
- Amazon EC2 Service Limits (Limites de serviço do Amazon EC2)
- What is Service Quotas? (O que é o Service Quotas?)
- How to Request Quota Increase (Como solicitar aumento de cota)
- Service endpoints and quotas (Endpoints e cotas de serviço)
- Guia do usuário do Service Quotas
- Quota Monitor for AWS (Monitor de cotas da AWS)
- AWS Fault Isolation Boundaries (Limites de isolamento de falhas da AWS)
- Availability with redundancy (Disponibilidade com redundância)
- AWS para dados
- O que significa integração contínua?
- O que significa distribuição contínua?
- Parceiro do APN: parceiros que podem ajudar no gerenciamento de configuração
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS (Gerenciar o ciclo devida da conta em ambientes de SaaS de conta por locatário na AWS)
- Managing and monitoring API throttling in your workloads (Gerenciar e monitorar o controle de utilização de API em workloads)
- View AWS Trusted Advisor recommendations at scale with AWS Organizations (Exibir recomendações do AWS Trusted Advisor em grande escala com AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower (Automatizar aumentos de limite de serviço e suporte empresarial com AWS Control Tower)

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Exibir e gerenciar cotas para serviços da AWS usando o Service Quotas)
- AWS IAM Quotas Demo (Demonstração de cotas do AWS IAM)

Serviços relacionados:

- Amazon CodeGuru Reviewer

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP03 Acomodar as restrições e as cotas fixas de serviço por meio da arquitetura

Esteja ciente das cotas de serviço, das restrições do serviço e dos limites de recursos físicos que não podem ser alterados. Projete arquiteturas para aplicações e serviços visando evitar que esses limites afetem a confiabilidade.

Os exemplos incluem largura de banda da rede, tamanho da carga útil da invocação da função sem servidor, taxa de intermitência de aceleração para um gateway da API e conexões simultâneas de usuários com um banco de dados.

Resultado desejado: a aplicação ou o serviço tem o desempenho esperado em condições de tráfego normal e alto. Elas foram projetadas para funcionar com as limitações referentes às restrições ficas ou cotas de serviço do recurso.

Antipadrões comuns:

- Escolher um design que usa um recurso de um serviço sem saber que há restrições de design que causarão falha à medida que você escala.
- Usar parâmetros de comparação irrealistas e que atingirão as cotas fixas do serviço durante os testes. Por exemplo, executar testes em um limite de intermitência mas por um período estendido.
- Escolher um design que não possa ser escalado nem modificado caso seja necessário ultrapassar as cotas fixas do serviço. Por exemplo, um tamanho de carga útil do SQS de 256 KB.
- A capacidade de observação não foi projetada nem implementada para monitorar e alertar sobre os limites das cotas de serviço que podem estar em risco durante eventos com tráfego alto.

Benefícios do estabelecimento dessa prática recomendada: verificar se a aplicação será executada em todos os níveis de carga de serviços projetados sem interrupção ou dano.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientações para a implementação

Ao contrário das cotas de serviço flexíveis ou de recursos que são substituídos com unidades de capacidade mais altas, as cotas fixas dos serviços da AWS não podem ser alteradas. Isso significa que todos esses tipos de serviços da AWS devem ser avaliados com relação a possíveis limites de capacidade rígidos quando usados em um design da aplicação.

Os limites rígidos são mostrados no console do Service Quotas. Se a coluna mostrar `ADJUSTABLE = No`, o serviço tem um limite rígido. Os limites rígidos também são mostrados em algumas páginas de configuração de recursos. Por exemplo, o Lambda tem limites rígidos específicos que não podem ser ajustados.

Como exemplo, ao projetar uma aplicação Python para ser executada em uma função do Lambda, a aplicação deve ser avaliada para determinar se há alguma chance de o Lambda ser executado por mais de 15 minutos. Se código puder ser executado mais do que esse limite de cota de serviço, tecnologias ou designs alterativos devem ser considerados. Se esse limite for atingido depois da implantação na produção, a aplicação sofrerá uma degradação e interrupção até que isso possa ser corrigido. Ao contrário das cotas flexíveis, não há um método para alterar esses limites mesmo sob eventos de emergência de gravidade 1.

Depois que a aplicação for implantada em um ambiente de teste, devem ser usadas estratégias para descobrir se algum limite rígido pode ser atingido. Testes de estresse, testes de carga e testes de caos devem fazer parte do plano de teste de introdução.

Etapas da implementação

- Revise a lista completa de serviços da AWS que poderiam ser usados na fase de design da aplicação.
- Revise os limites da cota flexível e os da cota rígida para todos esses serviços. Nem todos os limites são mostrados no console do Service Quotas. Alguns serviços descrevem esses limites em locais alternativos.
- À medida que você planeja a aplicação, revise os fatores que impulsionam a tecnologia e os negócios da workload, como resultados empresariais, casos de uso, sistemas dependentes, destinos de disponibilidade e objetos de recuperação de desastres. Permita que os fatores que

impulsionam a tecnologia e os negócios orientem o processo para identificar o sistema distribuído certo para sua workload.

- Analise a carga do serviço nas regiões e contas. Muitos limites rígidos são regionais para os serviços. No entanto, alguns limites são por conta.
- Analise arquiteturas de resiliência quanto ao uso de recursos durante uma falha de zona e de região. Na progressão de designs de várias regiões usando as abordagens ativo/ativo, ativo/passivo – quente, ativo/passivo – frio e ativo/passivo – luz-piloto, esses casos de falha resultarão em maior uso. Isso cria um possível caso de uso para atingir limites rígidos.

Recursos

Práticas recomendadas relacionadas:

- REL01-BP01 Conhecimento das cotas e restrições de serviço
- REL01-BP02 Gerenciar cotas de serviço de várias contas e regiões
- REL01-BP04 Monitorar e gerenciar cotas
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL01-BP06 Garantir que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover
- REL03-BP01 Escolher como segmentar a workload
- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência por meio da engenharia do caos

Documentos relacionados:

- AWS Pilar Confiabilidade da Well-Architected Framework: Disponibilidade
- AWS Service Quotas (anteriormente chamado de limites de serviço)
- AWS Trusted Advisor Best Practice Checks (Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Service Limits (Limites de serviço))
- AWS limit monitor on AWS answers (Monitor de limites da AWS em respostas da AWS)
- Amazon EC2 Service Limits (Limites de serviço do Amazon EC2)
- What is Service Quotas? (O que é o Service Quotas?)

- How to Request Quota Increase (Como solicitar aumento de cota)
- Service endpoints and quotas (Endpoints e cotas de serviço)
- Guia do usuário do Service Quotas
- Quota Monitor for AWS (Monitor de cotas da AWS)
- AWS Fault Isolation Boundaries (Limites de isolamento de falhas da AWS)
- Availability with redundancy (Disponibilidade com redundância)
- AWS para dados
- O que significa integração contínua?
- O que significa distribuição contínua?
- Parceiro do APN: parceiros que podem ajudar no gerenciamento de configuração
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS (Gerenciar o ciclo devida da conta em ambientes de SaaS de conta por locatário na AWS)
- Managing and monitoring API throttling in your workloads (Gerenciar e monitorar o controle de utilização de API em workloads)
- View AWS Trusted Advisor recommendations at scale with AWS Organizations (Exibir recomendações do AWS Trusted Advisor em grande escala com AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower (Automatizar aumentos de limite de serviço e suporte empresarial com AWS Control Tower)
- Ações, recursos e chaves de condição do Service Quotas

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Exibir e gerenciar cotas para serviços da AWS usando o Service Quotas)
- AWS IAM Quotas Demo (Demonstração de cotas do AWS IAM)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (AWS re:Invent 2018: fechar ciclos e abrir mentes: como controlar sistemas, sejam grandes ou pequenos)

Ferramentas relacionadas:

- AWS CodeDeploy

- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP04 Monitorar e gerenciar cotas

Avalie seu uso potencial e aumente suas cotas adequadamente, permitindo o crescimento planejado do uso.

Resultado desejado: sistemas ativos e automáticos que gerenciam e monitoram foram implantados. Essas soluções garantem que os limites de uso da cota sejam quase atingidos. Isso seria corrigido proativamente por mudanças na cota solicitada.

Antipadrões comuns:

- Não configurar o monitoramento para verificar os limites da cota de serviço.
- Não configurar o monitoramento de limites rígidos, embora esses valores não possam ser alterados.
- Presumir que o tempo necessário para solicitar e proteger uma mudança de cota flexível seja imediato ou período curto.
- Configurar alarmes para quando as cotas de serviço estiverem sendo atingidas, mas não ter um processo de resposta a um alerta.
- Configurar alarmes apenas para serviços compatíveis com o AWS Service Quotas e não monitorar outros serviços da AWS.
- Não considerar o gerenciamento da cota para designs com resiliência de várias regiões, como as abordagens ativo/ativo, ativo/passivo – quente, ativo/passivo – frio e ativo/passivo – luz-piloto.
- Não avaliar as diferenças de cota entre regiões.
- Não avaliar as necessidades de cada região com relação a uma solicitação de aumento de cota específica.

- Não utilizar modelos para o gerenciamento de cota de várias regiões.

Benefícios do estabelecimento dessa prática recomendada: o rastreamento automático do AWS Service Quotas e o monitoramento do uso em relação a essas cotas permitirão que você veja quando estiver perto de atingir um limite de cota. Também é possível usar esse monitoramento de dados para ajudar a limitar qualquer dano devido à exaustão da cota.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientações para a implementação

Para dispositivos compatíveis, é possível monitorar as cotas configurando vários serviços diferentes que podem avaliar e, então, enviar alertas ou alarmes. Isso pode auxiliar o monitoramento do uso e alertar quando você estiver se aproximando das cotas. Esses alarmes podem ser acionados pelo AWS Config, por funções do Lambda, pelo Amazon CloudWatch ou pelo AWS Trusted Advisor. Você também pode usar filtros de métrica no CloudWatch Logs para pesquisar e extrair padrões nos logs a fim de determinar se o uso está se aproximando dos limites de cota.

Etapas da implementação

Para monitoramento:

- Capture o consumo atual de recursos (por exemplo, buckets ou instâncias). Use operações de API de serviço, como a API do Amazon EC2 `DescribeInstances` para coletar o consumo atual de recursos.
- Capture as cotas atuais que são essenciais e aplicáveis aos serviços usando:
  - AWS Service Quotas
  - AWS Trusted Advisor
  - Documentação da AWS
  - Páginas específicas de serviços da AWS
  - AWS Command Line Interface (AWS CLI)
  - AWS Cloud Development Kit (AWS CDK)
- Use o AWS Service Quotas, um serviço da AWS que ajuda você a gerenciar as cotas de mais de 250 serviços da AWS em um único local.
- Use os limites de serviço do Trusted Advisor para monitorar os limites de serviço atuais em vários limites.
- Use o histórico da cota de serviço (console ou AWS CLI) para verificar os aumentos regionais.

- Compare as alterações na cota de serviço em cada região e cada conta para criar equivalência, se necessário.

Para gerenciamento:

- Automático: configure uma regra personalizada do AWS Config para verificar as cotas de serviço nas regiões e comparar as diferenças.
- Automático: configure uma função programada do Lambda para verificar as cotas de serviço nas regiões e comparar as diferenças.
- Manual: verifique as cotas de serviço por meio da AWS CLI, da API ou do Console da AWS para conferir as cotas de serviço nas regiões e comparar as diferenças. Relate as diferenças.
- Se forem identificadas diferenças nas cotas entre as regiões, solicite uma mudança na cota, se necessário.
- Avalie o resultado de todas as solicitações.

Recursos

Práticas recomendadas relacionadas:

- REL01-BP01 Conhecimento das cotas e restrições de serviço
- REL01-BP02 Gerenciar cotas de serviço de várias contas e regiões
- REL01-BP03 Acomodar as restrições e as cotas fixas de serviço por meio da arquitetura
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL01-BP06 Garantir que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover
- REL03-BP01 Escolher como segmentar a workload
- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência por meio da engenharia do caos

Documentos relacionados:

- AWS Pilar Confiabilidade da Well-Architected Framework: Disponibilidade

- AWS Service Quotas (anteriormente chamado de limites de serviço)
- AWS Trusted Advisor Best Practice Checks (Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Service Limits (Limites de serviço))
- AWS limit monitor on AWS answers (Monitor de limites da AWS em respostas da AWS)
- Amazon EC2 Service Limits (Limites de serviço do Amazon EC2)
- What is Service Quotas? (O que é o Service Quotas?)
- How to Request Quota Increase (Como solicitar aumento de cota)
- Service endpoints and quotas (Endpoints e cotas de serviço)
- Guia do usuário do Service Quotas
- Quota Monitor for AWS (Monitor de cotas da AWS)
- AWS Fault Isolation Boundaries (Limites de isolamento de falhas da AWS)
- Availability with redundancy (Disponibilidade com redundância)
- AWS para dados
- O que significa integração contínua?
- O que significa distribuição contínua?
- Parceiro do APN: parceiros que podem ajudar no gerenciamento de configuração
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS (Gerenciar o ciclo devida da conta em ambientes de SaaS de conta por locatário na AWS)
- Managing and monitoring API throttling in your workloads (Gerenciar e monitorar o controle de utilização de API em workloads)
- View AWS Trusted Advisor recommendations at scale with AWS Organizations (Exibir recomendações do AWS Trusted Advisor em grande escala com AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower (Automatizar aumentos de limite de serviço e suporte empresarial com AWS Control Tower)
- Ações, recursos e chaves de condição do Service Quotas

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Exibir e gerenciar cotas para serviços da AWS usando o Service Quotas)
- AWS IAM Quotas Demo (Demonstração de cotas do AWS IAM)

- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (AWS re:Invent 2018: fechar ciclos e abrir mentes: como controlar sistemas, sejam grandes ou pequenos)

Ferramentas relacionadas:

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

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 Garantir que existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover

Quando um recurso falha ou fica inacessível, ele ainda pode ser contabilizado nas cotas até ser encerrado com êxito. Verifique se as cotas abrangem a sobreposição de recursos inacessíveis ou com falha e suas substituições. Você deve considerar casos de uso como falha de rede, falha na zona de disponibilidade ou falhas regionais ao calcular essa lacuna.

Resultado desejado: falhas pequenas ou grandes em recursos ou na acessibilidade de recursos podem ser cobertas nos limites atuais do serviço. As falhas de zona, falhas de rede ou até mesmo falhas regionais têm sido consideradas no planejamento de recursos.

Antipadrões comuns:

- Configurar cotas de serviço com base nas necessidades atuais sem considerar os cenários de failover.
- Não considerar as entidades principais de estabilidade estática ao calcular a cota de pico de um serviço.
- Não considerar o potencial de recursos inacessíveis no cálculo da cota total necessária para cada região.
- Não considerar os limites de isolamento de falhas de serviço da AWS para alguns serviços e seus padrões de uso possivelmente anormais.

Benefícios do estabelecimento dessa prática recomendada: quando um evento de interrupção do serviço afeta a disponibilidade da aplicação, a nuvem permite implementar estratégias para mitigar ou se recuperar desses eventos. Essas estratégias geralmente incluem a criação de recursos adicionais para substituir os que falharam ou estão inacessíveis. Sua estratégia de cota acomodaria essas condições de failover e não incluiria danos adicionais devido à exaustão do limite de serviço.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientações para a implementação

Ao avaliar os limites de cota, considere casos de failover que podem ocorrer devido a algum dano. Os seguintes tipos de casos de failover devem ser considerados:

- Uma VPC interrompida ou inacessível.
- Uma sub-rede inacessível.
- Uma zona de disponibilidade foi danificada o suficiente para afetar a acessibilidade de muitos recursos.
- Várias rotas de rede ou pontos de ingresso e egresso são bloqueados ou alterados.
- Uma região foi danificada o suficiente para afetar a acessibilidade de muitos recursos.
- Há vários recursos, mas nem todos são afetados por uma falha em uma região ou zona de disponibilidade.

Falhas como as da lista acima poderiam ser o gatilho para iniciar um evento de failover. A decisão de fazer failover é única para cada situação e cliente, já que o impacto na empresa pode variar drasticamente. No entanto, ao decidir operacionalmente realizar failover de aplicações ou serviços, o planejamento da capacidade de recursos no local de failover e as cotas relacionadas devem ser solucionados antes do evento.

Revise as cotas de cada serviço considerando os picos mais altos do que o normal que podem ocorrer. Esses picos podem estar relacionados aos recursos que podem ser acessados devido às redes ou permissões, mas ainda estão ativos. Os recursos ativos não encerrados ainda serão contabilizados no limite de cota do serviço.

Etapas da implementação

- Verifique se há uma lacuna suficiente entre a cota de serviço e o uso máximo para acomodar um failover ou uma perda de acessibilidade.
- 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 o atendimento das solicitações de aumento de cota.
- Determine os requisitos de confiabilidade (também conhecidos como “número de noves”).

- Estabeleça seus 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 uma reserva adequada (por exemplo, 15%) do limite atual.
- Inclua cálculos para estabilidade estática (por zona e região), quando apropriado.
- Planeje o aumento do consumo (por exemplo, monitore suas tendências de consumo).
- Considere o impacto da estabilidade estática das suas workloads mais críticas. Avalie os recursos em conformidade com um sistema estaticamente estável em todas as regiões e zonas de disponibilidade.
- Considere o uso de reservas de capacidade sob demanda para programas a capacidade antecipadamente de qualquer failover. Isso pode ser uma estratégia útil durante as programações empresariais mais críticas para reduzir possíveis riscos de obter a quantidade o tipo certo de recursos durante o failover.

Recursos

Práticas recomendadas relacionadas:

- REL01-BP01 Conhecimento das cotas e restrições de serviço
- REL01-BP02 Gerenciar cotas de serviço de várias contas e regiões
- REL01-BP03 Acomodar as restrições e as cotas fixas de serviço por meio da arquitetura
- REL01-BP04 Monitorar e gerenciar cotas
- REL01-BP05 Automatizar o gerenciamento de cotas
- REL03-BP01 Escolher como segmentar a workload
- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência por meio da engenharia do caos

Documentos relacionados:

- AWS Pilar Confiabilidade da Well-Architected Framework: Disponibilidade

- AWS Service Quotas (anteriormente chamado de limites de serviço)
- AWS Trusted Advisor Best Practice Checks (Verificações de práticas recomendadas do AWS Trusted Advisor (consulte a seção Service Limits (Limites de serviço))
- AWS limit monitor on AWS answers (Monitor de limites da AWS em respostas da AWS)
- Amazon EC2 Service Limits (Limites de serviço do Amazon EC2)
- What is Service Quotas? (O que é o Service Quotas?)
- How to Request Quota Increase (Como solicitar aumento de cota)
- Service endpoints and quotas (Endpoints e cotas de serviço)
- Guia do usuário do Service Quotas
- Quota Monitor for AWS (Monitor de cotas da AWS)
- AWS Fault Isolation Boundaries (Limites de isolamento de falhas da AWS)
- Availability with redundancy (Disponibilidade com redundância)
- AWS para dados
- O que significa integração contínua?
- O que significa distribuição contínua?
- Parceiro do APN: parceiros que podem ajudar no gerenciamento de configuração
- Managing the account lifecycle in account-per-tenant SaaS environments on AWS (Gerenciar o ciclo devida da conta em ambientes de SaaS de conta por locatário na AWS)
- Managing and monitoring API throttling in your workloads (Gerenciar e monitorar o controle de utilização de API em workloads)
- View AWS Trusted Advisor recommendations at scale with AWS Organizations (Exibir recomendações do AWS Trusted Advisor em grande escala com AWS Organizations)
- Automating Service Limit Increases and Enterprise Support with AWS Control Tower (Automatizar aumentos de limite de serviço e suporte empresarial com AWS Control Tower)
- Ações, recursos e chaves de condição do Service Quotas

Vídeos relacionados:

- AWS Live re:Inforce 2019 - Service Quotas
- View and Manage Quotas for AWS Services Using Service Quotas (Exibir e gerenciar cotas para serviços da AWS usando o Service Quotas)
- AWS IAM Quotas Demo (Demonstração de cotas do AWS IAM)

- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (AWS re:Invent 2018: fechar ciclos e abrir mentes: como controlar sistemas, sejam grandes ou pequenos)

Ferramentas relacionadas:

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

CONFIABILIDADE 2. Como planejar sua topologia de rede?

Muitas vezes, as workloads 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

Criar uma conectividade de rede altamente disponível nos endpoints públicos das workloads pode ajudar a reduzir o tempo de inatividade devido à perda de conectividade e melhorar a disponibilidade e o SLA da workload. Para que isso seja possível, use DNS altamente disponível, redes de entrega de conteúdo (CDNs), gateways de API, balanceamento de carga ou proxies reversos.

Resultado desejado: é fundamental planejar, criar e operacionalizar a conectividade de rede altamente disponível para endpoints públicos. Se a workload ficar inacessível devido a uma perda de conectividade, mesmo se ela estiver em execução e indisponível, os clientes verão o sistema como inativo. Ao combinar a conectividade de rede altamente disponível e resiliente para os endpoints públicos da workload, junto com uma arquitetura resiliente para a própria workload, é possível fornecer o melhor nível possível de serviço e disponibilidade possível aos clientes.

O AWS Global Accelerator, o Amazon CloudFront, o Amazon API Gateway, os URLs de função do AWS Lambda, as APIs do AWS AppSync e o Elastic Load Balancing (ELB) fornecem endpoints públicos altamente disponíveis. O Amazon Route 53 fornece um serviço de DNS altamente disponível para a resolução do nome de domínio a fim de verificar se os endereços do endpoint público podem ser resolvidos.

Também é possível avaliar os dispositivos de software do AWS Marketplace com relação ao proxy e ao balanceamento de carga.

Antipadrões comuns:

- Projetar uma workload altamente disponível sem planejar a alta disponibilidade do DNS e da conectividade de rede.
- Usar endereços de internet públicos em instâncias ou contêineres individuais e gerenciar a conectividade com eles por meio de DNS.
- Usar endereços IP em vez de nomes de domínio para localizar serviços.
- Não testar cenários em que a conectividade com os endpoints públicos é perdida.
- Não analisar as necessidades de throughput de rede e os padrões de distribuição.
- Não testar nem se planejar para cenários em que a conectividade de rede da internet com os endpoints públicos da workload possam ser interrompidos.
- Fornecer conteúdo (como páginas da web, ativos estáticos ou arquivos de mídia) para uma grande área geográfica e não usar uma rede de entrega de conteúdo.
- Não se planejar para ataques de negação distribuída de serviços (DDoS). Ataques DDoS representam um risco de obstruir o tráfego legítimo e reduzir a disponibilidade para os usuários.

Benefícios do estabelecimento dessa prática recomendada: projetar-se para uma conectividade de rede resiliente e altamente disponível garante que a workload esteja acessível e disponível para os usuários.

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

Orientações para a implementação

No centro da criação de uma conectividade de rede altamente disponível com os endpoints públicos está o roteamento do tráfego. Para verificar se o tráfego consegue acessar os endpoints, o DNS deve poder resolver os nomes de domínio para os endereços IP correspondentes. Use um Sistema de Nomes de Domínio (DNS) escalável e altamente disponível, como o Amazon Route 53, para gerenciar os registros de DNS do domínio. Também é possível usar verificações de integridade fornecidas pelo Amazon Route 53. As verificações de integridade conferem se a aplicação está acessível, disponível e funcional, e podem ser configuradas de uma maneira que imitem o comportamento do usuário, como solicitar uma página da web ou um URL específico. Em caso de falha, o Amazon Route 53 responde às solicitações de resolução de DNS e direciona o tráfego somente aos endpoints íntegros. Também é possível considerar o uso dos recursos DNS GEO e Roteamento baseado em latência oferecidos pelo Amazon Route 53.

Para verificar se a própria workload é altamente disponível, use o Elastic Load Balancing (ELB). O Amazon Route 53 pode ser usado para direcionar o tráfego para o ELB, o que distribui o tráfego às instâncias de computação de destino. Também é possível usar o Amazon API Gateway com o AWS Lambda para obter uma solução sem servidor. Os clientes também podem executar workloads em várias Regiões da AWS. Com o padrão multissite ativo/ativo, a workload pode fornecer o tráfego de várias regiões. Com um padrão multissite ativo/passivo, a workload fornece tráfego da região ativa enquanto os dados são replicados para a região secundária e se torna ativa caso ocorra uma falha na região primaria. As verificações de integridade do Route 53 podem então ser usadas para controlar o failover de DNS de qualquer endpoint em uma região primária para um endpoint em uma região secundária, verificando se a workload está acessível e disponível para os usuários.

O Amazon CloudFront fornece uma API simples para distribuir o conteúdo com baixa latência e altas taxas de transferência de dados atendendo a solicitações usando uma rede de locais de borda ao redor do mundo. As redes de entrega de conteúdo (CDNs) atendem os clientes fornecendo conteúdo localizado ou armazenado em cache em um local próximo ao usuário. Isso também melhora a disponibilidade da aplicação, já que a carga do conteúdo é migrada dos servidores para os locais da borda do CloudFront. Os locais da borda e os caches de borda regionais armazenam cópias em cache do conteúdo próximo aos visualizadores, resultando em recuperação rápida e aumentando a acessibilidade e a disponibilidade da workload.

Para workloads com usuários distribuídos geograficamente, o AWS Global Accelerator ajuda a melhorar a disponibilidade e o desempenho das aplicações. O AWS Global Accelerator fornece endereços IP estáticos anycast que servem como um ponto de entrada fixo para a aplicação hospedada em uma ou mais Regiões da AWS. Isso permite que o tráfego entre na rede global da AWS o mais próximo possível dos usuários, melhorando a acessibilidade e a disponibilidade da workload. O AWS Global Accelerator também monitora a integridade dos endpoints da aplicação usando as verificações de integridade de TCP, HTTP e HTTPS. Qualquer mudança na integridade ou na configuração dos endpoints aciona o redirecionamento do tráfego de usuários para endpoints íntegros que oferecem o melhor desempenho e disponibilidade aos usuários. Além disso, o AWS Global Accelerator tem um design de isolamento de falhas que usa dois endereços IPv4 estáticos que são fornecidos por zonas de rede independentes, aumentando a disponibilidade das aplicações.

Para ajudar a proteger os clientes de ataques DDoS, a AWS oferece o AWS Shield Standard. O Shield Standard vem automaticamente habilitado e protege de ataques de infraestrutura comum (camadas três e quatro) como inundações de SYN/UDP e ataques de reflexão para comportar a alta disponibilidade das aplicações na AWS. Para obter mais proteções contra ataques maiores e mais sofisticados (como inundações de UDP), ataques de exaustão de estado (como inundações de TCP SYN) e para ajudar a proteger as aplicações executadas nos serviços Amazon Elastic Compute Cloud (Amazon EC2), Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator e Route 53, considere o uso do AWS Shield Advanced. Para proteção contra ataques de camada da aplicação como inundações de HTTP POST e GET, use o AWS WAF. O AWS WAF pode usar condições de scripts entre sites, endereços IP, cabeçalhos HTTP, corpo HTTP, strings de URI e injeção de SQL para determinar se uma solicitação deve ser bloqueada ou permitida.

Etapas da implementação

1. Configure DNS de alta disponibilidade: o Amazon Route 53 é um serviço da web de Sistema de Nomes de Domínio (DNS) escalável e altamente disponível. O Route 53 conecta as solicitações dos usuários às aplicações da internet executadas na AWS ou on-premises. Para obter mais informações, consulte Configurar o Amazon Route 53 como serviço DNS.
2. Configure verificações de integridade: ao usar o Route 53, verifique se somente os destinos íntegros podem ser resolvidos. Comece criando verificações de integridade do Route 53 e configurando o failover de DNS. Os aspectos a seguir são importantes a considerar ao configurar verificações de integridade:
    a. How Amazon Route 53 determines whether a health check is healthy (Como o Amazon Route 53 determina se uma verificação de integridade é íntegra)
    b. Criar, atualizar e excluir verificações de integridade

   c. Monitorar o status da verificação de integridade e receber notificações
   d. Práticas recomendadas do Amazon Route 53 DNS
3. Conectar o serviço de DNS aos endpoints.
   a. Ao usar o Elastic Load Balancing como destino do tráfego, crie um registro de alias usando o Amazon Route 53 que aponte para o endpoint regional do balanceador de carga. Durante a criação do registro de alias, defina a opção “Evaluate target health” (Avaliar integridade do destino) como “Yes” (Sim).
   b. Para workloads sem servidor ou APIs privadas quando o API Gateway é usado, use o Route 53 para redirecionar o tráfego para o API Gateway.
4. Decida sobre uma rede de entrega de conteúdo.
   a. Para entregar conteúdo usando locais da borda mais próximos ao usuário, comece entendendo como o CloudFront entrega conteúdo.
   b. Comece com uma distribuição simples do CloudFront. O CloudFront então sabe de onde você quer que o conteúdo seja entregue e os detalhes sobre como rastrear e gerenciar a entrega de conteúdo. É importante entender e considerar os aspectos a seguir ao configurar uma distribuição do CloudFront:
      i. Como funciona o armazenamento em cache com os pontos de presença do CloudFront
      ii. Aumentar a taxa de solicitações fornecidas diretamente de caches do CloudFront (taxa de acertos do cache)
      iii. Usar o Amazon CloudFront Origin Shield
      iv. Otimizar a alta disponibilidade com o failover de origem do CloudFront
5. Configure a proteção da camada da aplicação: o AWS WAF ajuda você a se proteger contra explorações e bots comuns da web que podem afetar a disponibilidade, comprometer a segurança ou consumir recursos em excesso. Para obter uma compreensão mais profunda, veja como o AWS WAF funciona e, quando estiver pronto para implementar proteções contra inundações HTTP POST E GET da camada de aplicações, consulte Getting started with AWS WAF (Conceitos básicos do AWS WAF). Também é possível usar o AWS WAF com o CloudFront. Consulte a documentação sobre como o AWS WAF funciona com os recursos do Amazon CloudFront.
6. Configure proteção adicional contra DDoS: por padrão, todos os clientes da AWS recebem proteção contra ataques de DDoS da camada de transporte e rede comuns e que ocorrem com mais frequência que visam seu site ou sua aplicação com o AWS Shield Standard sem custo adicional. Para proteção adicional de aplicações voltadas para a internet executadas no Amazon EC2, Elastic Load Balancing, Amazon CloudFront, AWS Global Accelerator e Amazon Route 53, considere o AWS Shield Advanced e consulte exemplos de arquiteturas resilientes a DDoS. Para

proteger sua workload e seus endpoints públicos de ataques de DDoS, consulte Getting started with AWS Shield Advanced (Conceitos básicos do AWS Shield Advanced).

Recursos

Práticas recomendadas relacionadas:

- REL10-BP01 Implantar a workload em vários locais
- REL10-BP02 Escolher os locais apropriados para sua implantação de vários locais
- REL11-BP04 Confiar no plano de dados e não no ambiente de gerenciamento durante a recuperação
- REL11-BP06 Enviar notificações quando os eventos afetarem a disponibilidade

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar a planejar sua rede
- AWS Marketplace para infraestrutura de rede
- O que é o Reachability Analyzer?
- O que é o Reachability Analyzer?
- O que é o Amazon Route 53?
- O que é o Reachability Analyzer?
- Network Connectivity capability - Establishing Your Cloud Foundations (Recurso de conectividade de rede: como estabelecer as bases da nuvem)
- O que é o Amazon API Gateway?
- What are AWS WAF, AWS Shield, and AWS Firewall Manager? (O que são o AWS WAF, o AWS Shield e o AWS Firewall Manager?)
- O que é o Amazon Route 53 Application Recovery Controller?
- Configurar verificações de integridade personalizadas para failover de DNS

Vídeos relacionados:

- AWS re:Invent 2022 - Improve performance and availability with AWS Global Accelerator(AWS re:Invent 2022: melhore o desempenho e a disponibilidade com o AWS Global Accelerator)

- AWS re:Invent 2020: Global traffic management with Amazon Route 53 (AWS re:Invent 2020: gerenciamento de tráfego global com o Amazon Route 53)
- AWS re:Invent 2022 - Operating highly available Multi-AZ applications (AWS re:Invent 2022: operar aplicações Multi-AZ altamente disponíveis)
- AWS re:Invent 2022 - Dive deep on AWS networking infrastructure (AWS re:Invent 2022: aprofundamento na infraestrutura de rede da AWS)
- AWS re:Invent 2022 - Building resilient networks (AWS re:Invent 2022: criar redes resilientes)

Exemplos relacionados:

- Disaster Recovery with Amazon Route 53 Application Recovery Controller (ARC) (Recuperação de desastres com o Amazon Route 53 Application Recovery Controller (ARC))
- Reliability Workshops (Workshops sobre confiabilidade)
- Workshop sobre o AWS Global Accelerator

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.

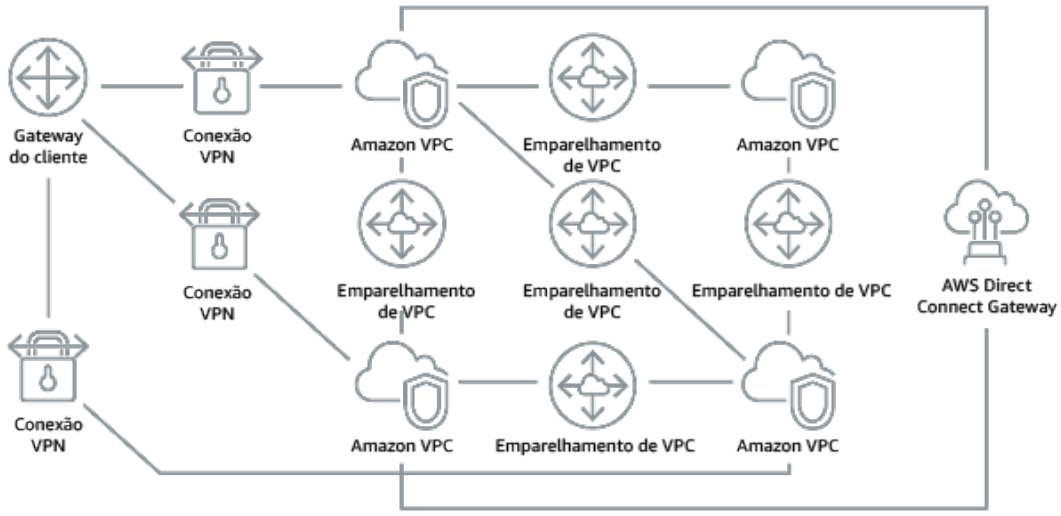


Figura 1: Sem o AWS Transit Gateway: você precisa emparelhar cada Amazon VPC com a outra e com cada localidade usando uma conexão VPN, que pode se tornar complexa à medida que ela escala.

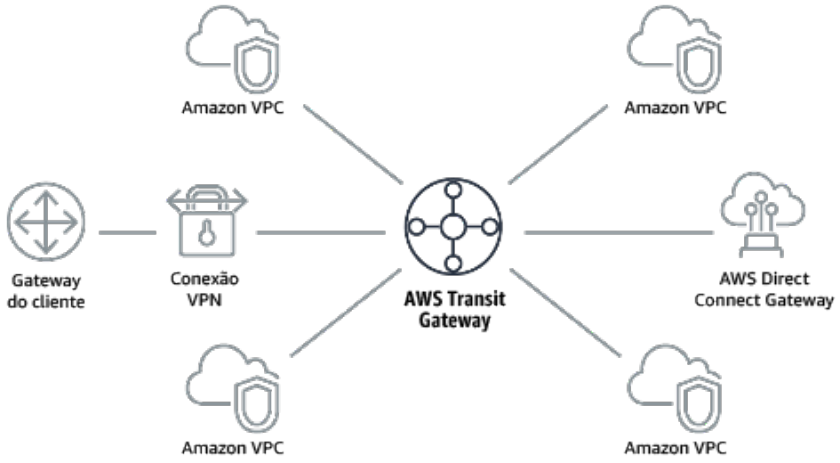


Figura 2: Com o AWS Transit Gateway: basta conectar cada Amazon VPC ou VPN ao AWS Transit Gateway e ele roteia o tráfego de e para cada VPC ou VPN.

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

- CONFIABILIDADE 3. Como projetar sua arquitetura de serviços de workload?
- CONFIABILIDADE 4. Como projetar interações em um sistema distribuído para evitar falhas?
- CONFIABILIDADE 5. Como projetar interações em um sistema distribuído para mitigar ou resistir a falhas?

CONFIABILIDADE 3. Como projetar sua arquitetura de serviços de workload?

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 enfocados em 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.

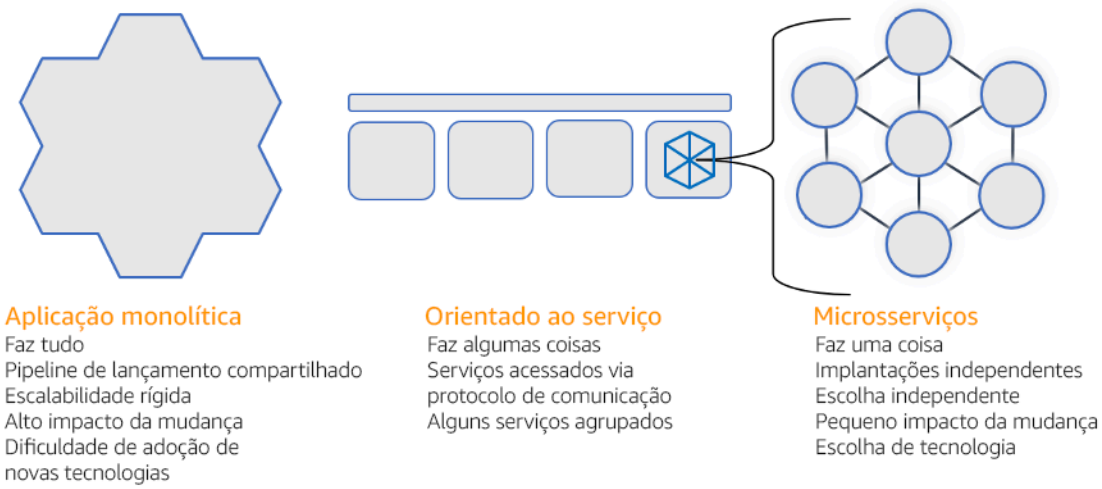
Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

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 enfocados em 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 enfocados em domínios e funcionalidades de negócios específicos

A arquitetura orientada a serviços (SOA) define serviços com funções bem delineadas que seguem as necessidades dos negócios. Os microsserviços usam modelos de domínio e contexto delimitado para traçar limites de serviço ao longo dos limites do contexto de negócios. O foco nos domínios de negócios e na funcionalidade ajuda as equipes a definir requisitos independentes de confiabilidade para seus serviços. Contextos delimitados isolam e encapsulam a lógica de negócios, permitindo que as equipes raciocinem melhor sobre como lidar com falhas.

Resultado desejado: Em conjunto, engenheiros e partes interessadas do negócio definem contextos delimitados e os usam para projetar sistemas como serviços que cumprem funções empresariais específicas. Essas equipes usam práticas estabelecidas, como Event Storming, para definir os requisitos. As novas aplicações são projetadas como serviços, limites bem definidos e acoplamento fraco. Os monólitos existentes são decompostos em contextos delimitados e os projetos de sistemas migram para arquiteturas SOA ou de microsserviços. Quando os monólitos são refatorados, abordagens estabelecidas, como contextos de bolha e padrões de decomposição de monólitos, são aplicadas.

Os serviços orientados a domínios são executados como um ou mais processos que não compartilham o estado. Eles respondem de forma independente às flutuações na demanda e lidam com cenários de falha à luz dos requisitos específicos do domínio.

Antipadrões comuns:

- As equipes são formadas em torno de domínios técnicos específicos, como UI e UX, middleware ou banco de dados, em vez de domínios empresariais específicos.
- As aplicações abrangem as responsabilidades do domínio. Serviços que abrangem contextos delimitados podem ser mais difíceis de manter, exigir maiores esforços de teste e que várias equipes de domínio participem das atualizações de software.
- As dependências de domínio, como as bibliotecas de entidades de domínio, são compartilhadas entre serviços, de forma que as alterações em um domínio de serviço exijam alterações em outros domínios de serviço.
- Os contratos de serviço e a lógica de negócios não expressam entidades em uma linguagem de domínio comum e consistente, ocasionando camadas de tradução que complicam os sistemas e aumentam os esforços de depuração.

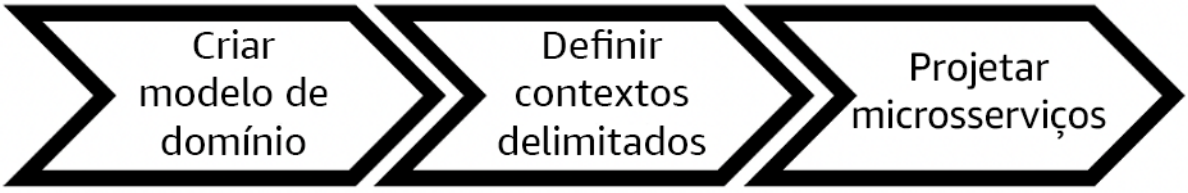
Benefícios de estabelecer esta prática recomendada: As aplicações são projetadas como serviços independentes delimitados por domínios de negócios e usam uma linguagem comercial comum.

Os serviços podem ser testados e implantados de forma independente. Os serviços atendem aos requisitos de resiliência específicos do domínio implementado.

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

Orientação para implementação

A decisão orientada por domínio (DDD) é a abordagem fundamental para projetar e criar software em torno de domínios empresariais. É útil trabalhar com uma framework existente ao criar serviços enfocados em domínios empresariais. Ao trabalhar com aplicações monolíticas existentes, você pode utilizar os padrões de decomposição que fornecem técnicas estabelecidas para modernizar aplicações em serviços.



Decisão orientada por domínio

Etapas da implementação

- As equipes podem realizar workshops de Event Storming a fim de identificar rapidamente eventos, comandos, agregados e domínios em um formato leve de notas adesivas.
- Depois que as entidades e as funções do domínio forem formadas em um contexto de domínio, você poderá dividir seu domínio em serviços usando contexto delimitado, em que entidades que compartilham recursos e atributos semelhantes são agrupadas. Com o modelo dividido em contextos, surge um modelo de como delimitar microsserviços.
  - Por exemplo, as entidades do site Amazon.com podem incluir pacote, entrega, programação, preço, desconto e moeda.
  - Pacote, entrega e cronograma são agrupados no contexto de envio, enquanto preço, desconto e moeda são agrupados no contexto de preços.
- Decompor monólitos em microsserviços descreve padrões para refatorar microsserviços. O uso de padrões para decomposição por capacidade comercial, subdomínio ou transação se alinha bem às abordagens orientadas por domínio.
- Técnicas táticas, como o contexto de bolha, permitem introduzir o DDD em aplicações existentes ou legadas sem reformulações antecipadas e compromissos totais com o DDD. Em uma

abordagem de contexto de bolha, um pequeno contexto delimitado é estabelecido usando um mapeamento e coordenação de serviços, ou camada anticorrupção, que protege o modelo de domínio recém-definido de influências externas.

Depois que as equipes realizarem a análise de domínio e definirem entidades e contratos de serviço, elas podem utilizar os serviços da AWS para implementar o design orientado por domínio como serviços baseados em nuvem.

- Comece o desenvolvimento definindo testes que exercitem as regras de negócios de seu domínio. O desenvolvimento orientado por testes (TDD) e o desenvolvimento orientado por comportamento (BDD) ajudam as equipes a manter os serviços enfocados na solução de problemas de negócios.
- Selecione os serviços da AWS que mais bem atendam aos requisitos de domínio de sua empresa e arquitetura de microsserviços:
  - tecnologia sem servidor da AWS permite que sua equipe enfoque a lógica de domínio específica em vez de gerenciar servidores e infraestrutura.
  - Contêineres na AWS simplificam o gerenciamento de sua infraestrutura para que você possa enfocar nos requisitos de domínio.
  - Bancos de dados com propósito específico ajudam você a adequar seus requisitos de domínio ao tipo de banco de dados mais adequado.
- Criação de arquiteturas hexagonais na AWS descreve uma framework para criar lógica de negócios em serviços que funcionam retroativamente a partir de um domínio empresarial para atender aos requisitos funcionais e, depois, conectar adaptadores de integração. Os padrões que separam os detalhes da interface da lógica de negócios com serviços da AWS ajudam as equipes a enfocar na funcionalidade do domínio e melhorar a qualidade do software.

Recursos

Práticas recomendadas relacionadas:

- REL03-BP01 Escolher como segmentar a workload
- REL03-BP03 Fornecer contratos de serviço por API

Documentos relacionados:

- AWS Microsserviços
- Implementação de microsserviços na AWS

- How to break a Monolith into Microservices (Como dividir um monólito em microsserviços)
- Getting Started with DDD when Surrounded by Legacy Systems (Conceitos básicos do DDD quando cercado por sistemas herdados)
- Domain-Driven Design: Tackling Complexity in the Heart of Software (Design orientado por domínio: como lidar com a complexidade no núcleo do software)
- Criação de arquiteturas hexagonais na AWS
- Decompor monólitos em microsserviços
- Event Storming
- Mensagens entre contextos delimitados
- Microsserviços
- Desenvolvimento orientado por testes
- Desenvolvimento orientado pelo comportamento

Exemplos relacionados:

- Workshop nativo da nuvem corporativa
- Designing Cloud Native Microservices on AWS (from DDD/EventStormingWorkshop) (Como projetar microsserviços nativos em nuvem na AWS (do DDD/EventStormingWorkshop))

Ferramentas relacionadas:

- Bancos de dados da Nuvem AWS
- Tecnologia sem servidor na AWS
- Contêineres na AWS

REL03-BP03 Fornecer contratos de serviço por API

Os contratos de serviço são acordos documentados entre produtores e consumidores de API estabelecidos em uma definição de API legível por máquina. Uma estratégia de versionamento de contrato permite que os consumidores continuem usando a API existente e migrem suas aplicações para uma API mais recente quando estiverem prontos. A implantação do produtor pode acontecer a qualquer momento, desde que o contrato seja cumprido. A equipe de serviços pode usar a pilha de tecnologia de sua preferência para cumprir o contrato de API.

Resultado desejado:

Antipadrões comuns: As aplicações criadas com arquiteturas orientadas a serviços ou microsserviços podem operar de forma independente e, ao mesmo tempo, ter uma dependência de runtime integrada. As alterações implantadas em um consumidor ou produtor de API não interrompem a estabilidade do sistema geral quando os dois lados seguem um contrato de API comum. Os componentes que se comunicam por meio de APIs de serviço podem realizar lançamentos funcionais independentes, atualizações para dependências de runtime ou fazer failover em um site de recuperação de desastres (DR) com pouco ou nenhum impacto entre si. Além disso, serviços diferentes são capazes de escalar de forma independente a absorção da demanda de recursos sem exigir que outros serviços escalem simultaneamente.

- Criar APIs de serviço sem esquemas altamente tipificados. Isso ocasiona APIs que não podem ser usadas para gerar vinculações de API e payloads que não possam ser validadas de maneira programática.
- Não adotar uma estratégia de versionamento, o que força os consumidores de API a atualizarem e lançarem ou falharem com a evolução dos contratos de serviço.
- Mensagens de erro que vazam detalhes da implementação do serviço subjacente em vez de descreverem falhas de integração no contexto e no idioma do domínio.
- Não usar contratos de API para desenvolver casos de teste e simular implementações de API para permitir testes independentes dos componentes do serviço.

Benefícios de estabelecer esta prática recomendada: Sistemas distribuídos compostos por componentes que se comunicam por meio de contratos de serviço de API podem aumentar a confiabilidade. Os desenvolvedores podem detectar possíveis problemas no início do processo de desenvolvimento com a verificação de tipo durante a compilação a fim de verificar se as solicitações e as respostas seguem o contrato da API e se os campos obrigatórios estão presentes. Os contratos de API oferecem uma interface clara de autodocumentação de APIs e oferecem melhor interoperabilidade entre diferentes sistemas e linguagens de programação.

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

Orientação para implementação

Depois de identificar os domínios de negócios e determinar a segmentação da workload, você pode desenvolver suas APIs de serviço. Primeiro, defina contratos de serviço legíveis por máquina para APIs e, depois, implemente uma estratégia de versionamento de API. Quando estiver pronto para

integrar serviços em protocolos comuns, como REST, GraphQL ou eventos assíncronos, você poderá incorporar serviços da AWS à sua arquitetura para integrar seus componentes com contratos de API altamente tipificados.

Serviços da AWS para contratos de API de serviços

Incorpore serviços da AWS, incluindo Amazon API Gateway, o AWS AppSync e o Amazon EventBridge à sua arquitetura para usar contratos de serviço de API em sua aplicação. O Amazon API Gateway ajuda você a se integrar diretamente a serviços nativos da AWS e outros serviços da web. O API Gateway é compatível com a especificação da OpenAPI e versionamento. O AWS AppSync é um endpoint GraphQL gerenciado que você configura definindo um esquema GraphQL para definir uma interface de serviço para consultas, mutações e assinaturas. O Amazon EventBridge usa esquemas de eventos para definir eventos e gerar associações de código para seus eventos.

Etapas da implementação

- Primeiro, defina um contrato para sua API. Um contrato expressará os recursos de uma API, bem como definirá objetos e campos de dados altamente tipificados para a entrada e a saída da API.
- Ao configurar APIs no API Gateway, você pode importar e exportar especificações da OpenAPI para seus endpoints.
  - A importação de uma definição de OpenAPI simplifica a criação de sua API e pode ser integrada a ferramentas de infraestrutura como código da AWS, como o AWS Serverless Application Model e o AWS Cloud Development Kit (AWS CDK).
  - A exportação de uma definição de API simplifica a integração a ferramentas de teste de API e oferece ao consumidor de serviços uma especificação de integração.
- Você pode definir e gerenciar APIs do GraphQL com o AWS AppSync definindo um arquivo de esquema GraphQL para gerar sua interface de contrato e simplificar a interação com modelos REST complexos, várias tabelas de banco de dados ou serviços legados.
- Projetos do AWS Amplify integrados ao AWS AppSync geram arquivos de consulta JavaScript altamente tipificados para uso em sua aplicação, bem como uma biblioteca cliente do AWS AppSync GraphQL para tabelas do Amazon DynamoDB .
- Quando você consome eventos de serviço do Amazon EventBridge, eles seguem os esquemas já existentes no registro do esquema ou os definidos com a especificação da OpenAPI. Com um esquema definido no registro, também é possível gerar vinculações de cliente a partir do contrato de esquema para integrar seu código aos eventos.

- Estender ou realizar o versionamento de sua API. Estender uma API é uma opção mais simples ao adicionar campos que podem ser configurados com campos opcionais ou valores padrão para campos obrigatórios.
  - Contratos baseados em JSON para protocolos, como REST e GraphQL, podem ser uma boa opção para a extensão do contrato.
  - Contratos baseados em XML para protocolos, como SOAP, devem ser testados com consumidores de serviços para determinar a viabilidade da extensão do contrato.
- Ao realizar o versionamento de uma API, considere implementar o controle de versão por procuração em que uma fachada é usada para oferecer compatibilidade com versões para que a lógica possa ser mantida em uma única base de código.
  - Com o API Gateway, você pode usar mapeamentos de solicitações e respostas para simplificar a absorção de alterações no contrato estabelecendo uma fachada para fornecer valores padrão para novos campos ou para retirar os campos removidos de uma solicitação ou resposta. Com essa abordagem, o serviço subjacente pode manter uma única base de código.

Recursos

Práticas recomendadas relacionadas:

- REL03-BP01 Escolher como segmentar a workload
- REL03-BP02 Criar serviços enfocados em domínios e funcionalidades de negócios específicos
- REL04-BP02 Implementar dependências com acoplamento fraco
- REL05-BP03 Controlar e limitar as chamadas de repetição
- REL05-BP05 Definir tempos limite do cliente

Documentos relacionados:

- O que é uma API (interface de programação de aplicações)?
- Implementação de microsserviços na AWS
- Compensações de microsserviços
- Microsserviços - uma definição desse novo termo de arquitetura
- Microsserviços na AWS
- Trabalhar com extensões do API Gateway para OpenAPI
- Especificação da OpenAPI

- GraphQL: esquemas e tipos
- Vinculações de código do Amazon EventBridge

Exemplos relacionados:

- Amazon API Gateway: configurar uma API REST usando o OpenAPI
- Amazon API Gateway para a aplicação CRUD Amazon DynamoDB usando OpenAPI
- Padrões modernos de integração de aplicações em uma era sem servidor: integração do serviço API Gateway
- Implementar o versionamento baseado em cabeçalho do API Gateway com Amazon CloudFront
- AWS AppSync: criar uma aplicação cliente

Vídeos relacionados:

- Usar a OpenAPI no AWS SAM para gerenciar o API Gateway

Ferramentas relacionadas:

- Amazon API Gateway
- AWS AppSync
- Amazon EventBridge

CONFIABILIDADE 4. Como projetar 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.

Em sistemas fortemente acoplados, as mudanças em um componente podem exigir mudanças em outros componentes que dependem dele, o que resulta em desempenho degradado em todos eles. O acoplamento fraco interrompe essa dependência de forma que os componentes dependentes só precisem conhecer 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 acoplamento fraco permite modificar o código ou adicionar recursos a um componente, minimizando o risco para outros componentes que dependem dele. Ele também permite resiliência detalhada em nível de componente, caso em que é possível aumentar a escala horizontalmente ou até mesmo alterar a implementação subjacente 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 Amazon SQS, uma plataforma de dados de streaming, como o AWS Step Functions, ou o Amazon Kinesis.

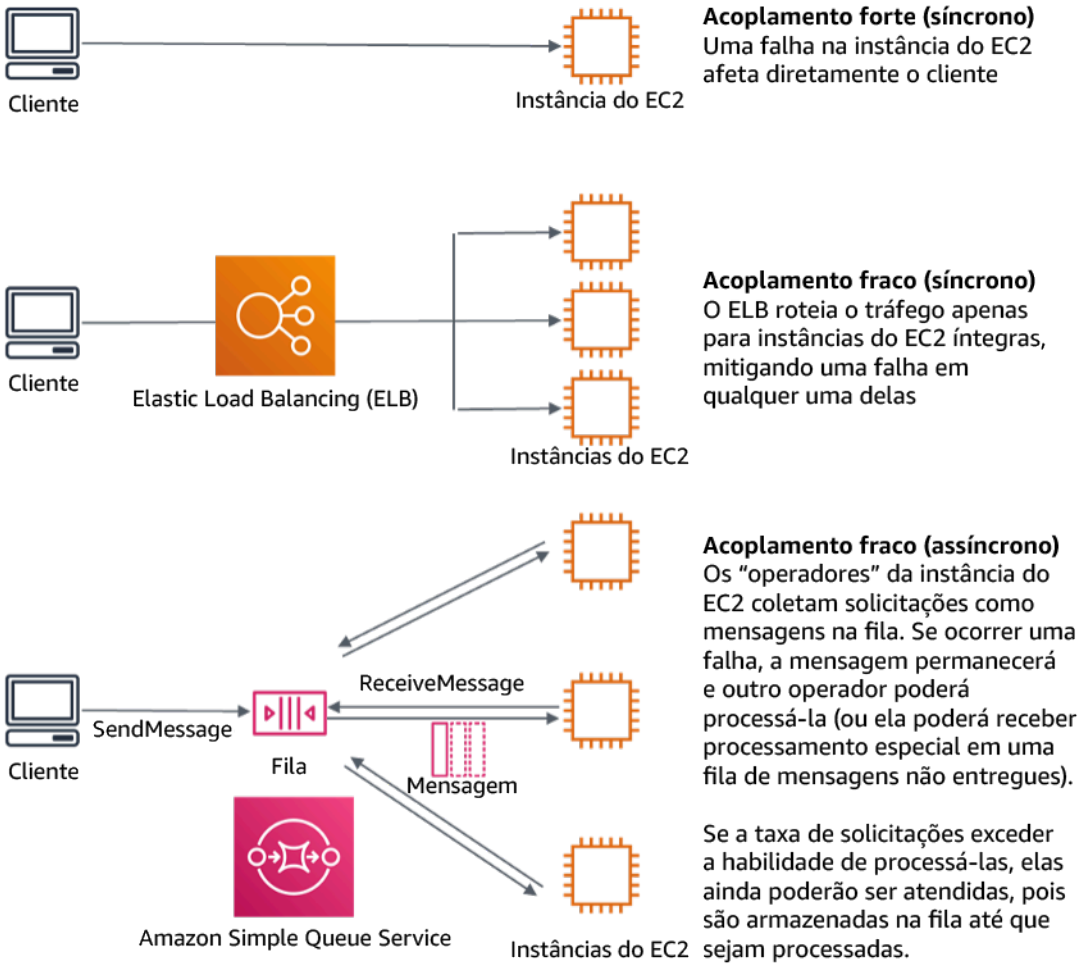


Figura 4: Dependências como sistemas de enfileiramento e load balancers têm baixo acoplamento

Filas do Amazon SQS e Elastic Load Balancers são apenas duas maneiras de adicionar uma camada intermediária para baixo acoplamento. Arquiteturas orientadas a 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.

Resultado desejado: a implementação de dependências com acoplamento fraco permite minimizar a área de superfície da falha para o nível de componente, o que ajuda a diagnosticar e resolver problemas. Também simplifica os ciclos de desenvolvimento, permitindo que as equipes implementem mudanças em um nível modular sem impactar o desempenho de outros componentes que dependem delas. Essa abordagem fornece a capacidade de aumentar a escala horizontalmente em nível de componente com base nas necessidades dos recursos, bem como na utilização de um componente que contribui para a redução de custos.

Antipadrões comuns:

- Implantar uma workload monolítica.
- Invocar diretamente as APIs entre níveis de workload sem recurso de failover ou processamento assíncrono da solicitação.
- Acoplamento forte usando dados compartilhados. Sistemas com acoplamento fraco devem evitar o compartilhamento de dados por meio de bancos de dados compartilhados ou outras formas de armazenamento de dados fortemente acoplado, o que pode reintroduzir o acoplamento forte e impedir a escalabilidade.
- Ignorar a contrapressão. A workload deve ter a capacidade de diminuir ou interromper a entrada de dados quando um componente não puder processá-los na mesma velocidade.

Benefícios do estabelecimento dessa prática recomendada: o acoplamento fraco 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 se esta prática recomendada não for estabelecida: alto

Orientações para a implementação

Implemente dependências com acoplamento fraco. Existem várias soluções que permitem criar aplicações com acoplamento fraco. Isso inclui serviços para implementar filas totalmente gerenciadas, fluxos de trabalho automatizados, reação a eventos e APIs, entre outros, que podem ajudar a isolar o comportamento de componentes de outros componentes e, dessa forma, aumentar a resiliência e a agilidade.

- Criar arquiteturas orientadas a eventos: o Amazon EventBridge ajuda a criar arquiteturas orientadas a eventos com acoplamento fraco e distribuídas.
- Implementar filas em sistemas distribuídos: é possível usar o Amazon Simple Queue Service (Amazon SQS) para integrar e desacoplar sistemas distribuídos.

- Conteinerizar componentes como microsserviços: os microsserviços permitem que as equipes criem aplicações constituídas de pequenos componentes independentes que se comunicam por meio de APIs bem definidas. O Amazon Elastic Container Service (Amazon ECS) e o Amazon Elastic Kubernetes Service (Amazon EKS) podem ajudar você a começar a usar contêineres mais rapidamente.
- Gerenciar fluxos de trabalho com o Step Functions: o Step Functions ajuda você a coordenar vários serviços da AWS em fluxos de trabalho flexíveis.
- Utilizar as arquiteturas de mensagens de publicação e assinatura (pub/sub): o Amazon Simple Notification Service (Amazon SNS) fornece a entrega de mensagens dos publicadores aos assinantes (também conhecidos como produtores e consumidores).

Etapas da implementação

- Os componentes em uma arquitetura orientada a eventos são iniciados por eventos. Eventos são ações que ocorrem em um sistema, como um usuário que adiciona um item a um carrinho. Quando uma ação é bem-sucedida, é gerado um evento que aciona o próximo componente do sistema.
  - Building Event-driven Applications with Amazon EventBridge
  - AWS re:Invent 2022 - Designing Event-Driven Integrations using Amazon EventBridge
- Os sistemas de mensagens distribuídos têm três partes principais que precisam ser implementadas para uma arquitetura baseada em fila. Eles incluem componentes do sistema distribuído, a fila usada para desacoplamento (distribuída em servidores do Amazon SQS) e as mensagens na fila. Um sistema típico tem produtores que iniciam a mensagem na fila e o consumidor que recebe a mensagem da fila. A fila armazena as mensagens em vários servidores do Amazon SQS para redundância.
  - Basic Amazon SQS architecture
  - Send Messages Between Distributed Applications with Amazon Simple Queue Service
- Os microsserviços, quando bem utilizados, melhoram a capacidade de manutenção e aumentam a escalabilidade, pois os componentes com acoplamento fraco são gerenciados por equipes independentes. Também permitem o isolamento de comportamentos em um único componente em caso de mudanças.
  - Implementação de microsserviços na AWS
  - Let's Architect! Architecting microservices with containers

- Com o AWS Step Functions é possível criar aplicações distribuídas, automatizar processos, orquestrar microsserviços, entre outras coisas. A orquestração de vários componentes em um fluxo de trabalho automatizado permite desacoplar as dependências na aplicação.
  - Create a Serverless Workflow with AWS Step Functions and AWS Lambda
  - Conceitos básicos do AWS Step Functions

Recursos

Documentos relacionados:

- Amazon EC2: Ensuring Idempotency
- 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?
- Break up with your monolith
- Orchestrate Queue-based Microservices with AWS Step Functions and Amazon SQS
- Basic Amazon SQS architecture
- Arquitetura baseada em fila

Vídeos relacionados:

- AWS New York Summit 2019: Introduction 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 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)
- AWS re:Invent 2019: Scalable serverless event-driven applications using Amazon SQS and Lambda
- AWS re:Invent 2022 - Designing event-driven integrations using Amazon EventBridge
- AWS re:Invent 2017: Elastic Load Balancing Deep Dive and Best Practices

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)

CONFIABILIDADE 5. Como projetar 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 workload. Essas práticas recomendadas permitem que as workloads 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 para recuperação) é melhorado.

Práticas recomendadas

- REL05-BP01 Implementar uma degradação simples para transformar dependências rígidas aplicáveis em dependências flexíveis
- REL05-BP02 Controlar a utilização 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

Os componentes da aplicação devem continuar desempenhando sua função principal mesmo que as dependências se tornem indisponíveis. Eles podem estar fornecendo dados um pouco obsoletos, dados alternativos ou até mesmo nenhum dado. Isso garante que o funcionamento geral do sistema seja minimamente impedido por falhas localizadas e, ao mesmo tempo, ofereça o valor empresarial central.

Resultado desejado: Quando as dependências de um componente não estão íntegras, o próprio componente ainda pode funcionar, embora de maneira prejudicada. Os modos de falha dos componentes devem ser vistos como operação normal. Os fluxos de trabalho devem ser projetados de forma que essas falhas não ocasionem à falha total ou, pelo menos, a estados previsíveis e recuperáveis.

Antipadrões comuns:

- Não identificar a principal funcionalidade empresarial necessária. Não testar se os componentes estão funcionando mesmo durante falhas de dependência.
- Não fornecer dados sobre erros ou quando apenas uma das várias dependências não está disponível e resultados parciais ainda podem ser retornados.
- Criar um estado inconsistente quando uma transação falha parcialmente.
- Não ter uma forma alternativa de acessar um armazenamento de parâmetros central.

- Invalidar ou esvaziar o estado local como resultado de uma falha na atualização sem levar em conta as consequências de fazer isso.

Benefícios de estabelecer esta prática recomendada: A degradação gradual melhora a disponibilidade do sistema como um todo e mantém as funções mais importantes em execução mesmo durante falhas.

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

Orientação para implementação

A implementação de uma degradação gradual ajuda a minimizar o impacto das falhas de dependência na função do componente. Preferencialmente, um componente detecta falhas de dependência e as contorna de uma maneira que afete minimamente outros componentes ou clientes.

Arquitetar para uma degradação gradual significa considerar possíveis modos de falha durante o projeto de dependência. Para cada modo de falha, tenha uma maneira de fornecer a maior parte ou pelo menos a funcionalidade mais crítica do componente para chamadores ou clientes. Essas considerações podem se tornar requisitos adicionais que podem ser testados e verificados. Preferencialmente, um componente é capaz de realizar sua função principal de maneira aceitável, mesmo quando uma ou várias dependências falhem.

Trata-se tanto de uma discussão empresarial quanto técnica. Todos os requisitos empresariais são importantes e devem ser atendidos, se possível. No entanto, ainda faz sentido perguntar o que deve acontecer quando nem todos eles podem ser cumpridos. Um sistema pode ser projetado para estar disponível e ser consistente, mas em circunstâncias em que um requisito deve ser descartado, qual deles é mais importante? Para o processamento de pagamentos, pode ser a consistência. Para uma aplicação em tempo real, pode ser a disponibilidade. Para um site voltado para o cliente, a resposta pode depender das expectativas do cliente.

O que isso significa depende dos requisitos do componente e do que deve ser considerado sua função principal. Por exemplo:

- Um site de comércio eletrônico pode exibir dados de vários sistemas diferentes, como recomendações personalizadas, produtos mais bem classificados e status dos pedidos dos clientes na página de pouso. Quando um sistema upstream falha, ainda faz sentido exibir todo o resto em vez de mostrar uma página de erro para um cliente.
- Um componente que executa gravações em lote ainda poderá continuar processando um lote se ocorrer uma falha em uma das operações individuais. Deve ser simples implementar um

mecanismo de repetição. Isso pode ser feito retornando informações sobre quais operações foram bem-sucedidas, quais falharam e por que falharam para o chamador, ou colocando solicitações com falha em uma fila de mensagens não entregues para implementar repetições assíncronas. As informações sobre operações com falha também devem ser registradas em log.

- Um sistema que processa transações deve verificar se todas ou nenhuma atualização individual foi executada. Para transações distribuídas, o padrão saga pode ser usado para reverter operações anteriores caso ocorra uma falha em uma operação posterior da mesma transação. Aqui, a função principal é manter a consistência.
- Sistemas essenciais devem ser capazes de lidar com dependências não correspondentes em tempo hábil. Nesses casos, o padrão do disjuntor pode ser usado. Quando as respostas de uma dependência começam a atingir o tempo limite, o sistema pode mudar para um estado fechado em que nenhuma chamada adicional é realizada.
- Uma aplicação pode ler parâmetros de um armazenamento de parâmetros. Pode ser útil criar imagens de contêiner com um conjunto padrão de parâmetros e usá-las caso o armazenamento de parâmetros não esteja disponível.

Observe que as vias percorridas em caso de falha do componente precisam ser testadas e devem ser significativamente mais simples do que a via principal. Geralmente, estratégias alternativas devem ser evitadas.

Etapas da implementação

Identifique dependências externas e internas. Leve em conta quais tipos de falhas podem ocorrer nelas. Pense em maneiras de minimizar o impacto negativo nos sistemas upstream e downstream e nos clientes durante essas falhas.

Veja a seguir uma lista de dependências e como degradar normalmente quando elas falham:

1. Falha parcial das dependências: Um componente pode fazer várias solicitações para sistemas downstream, como várias solicitações para um sistema ou uma solicitação para vários sistemas cada. Dependendo do contexto empresarial, diferentes maneiras de lidar com isso podem ser apropriadas (para obter mais detalhes, consulte exemplos anteriores em Orientações de implementação).
2. Um sistema downstream não consegue processar solicitações devido à alta carga: Se as solicitações para um sistema downstream falharem de modo consistente, não faz sentido continuar tentando. Isso pode criar carga adicional em um sistema já sobrecarregado e dificultar a recuperação. O padrão do disjuntor pode ser utilizado aqui, que monitora as chamadas com

falha para um sistema downstream. Se ocorrer uma falha em um grande número de chamadas, ele deixará de enviar mais solicitações para o sistema downstream e só ocasionalmente permitirá que as chamadas passem para testar se o sistema downstream está disponível novamente.

3. Um armazenamento de parâmetros não está disponível: Para transformar um armazenamento de parâmetros, é possível usar o armazenamento em cache flexível de dependências ou padrões razoáveis incluídos nas imagens do contêiner ou da máquina. Observe que esses padrões precisam ser mantidos atualizados e incluídos nos pacotes de testes.
4. Um serviço de monitoramento ou outra dependência não funcional não está disponível: Se um componente não conseguir enviar logs, métricas ou rastreamentos de forma intermitente para um serviço de monitoramento central, geralmente é melhor continuar executando as funções empresariais normalmente. Não registrar em log nem enviar métricas silenciosamente por um longo período geralmente não é aceitável. Além disso, alguns casos de uso podem exigir entradas de auditoria completas para atender aos requisitos de conformidade.
5. Uma instância primária de um banco de dados relacional pode não estar disponível: Amazon Relational Database Service, como quase todos os bancos de dados relacionais, só pode ter uma instância primária de gravador. Isso cria um único ponto de falha para workloads de gravação e dificulta o ajuste de escala. Isso pode ser parcialmente reduzido com o uso de uma configuração Multi-AZ para alta disponibilidade ou da tecnologia sem servidor da Amazon Aurora para melhor ajuste de escala. Para requisitos de disponibilidade muito altos, pode fazer sentido não confiar no gravador principal. Para consultas que são somente leitura, podem ser usadas réplicas de leitura, que fornecem redundância e a capacidade de aumentar a escala horizontalmente, não apenas para verticalmente. As gravações podem ser armazenadas em buffer, por exemplo, em uma fila do Amazon Simple Queue Service, para que as solicitações de gravação dos clientes ainda possam ser aceitas mesmo que a primária esteja temporariamente indisponível.

Recursos

Documentos relacionados:

- Amazon API Gateway: controlar a utilização das solicitações de API para um melhor throughput
- 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) (Repetição, recuo e jitter: AWS re:Invent 2019: Introdução à 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 a utilização de solicitações

Controle a utilização das solicitações para reduzir o esgotamento de recursos devido a aumentos inesperados na demanda. Solicitações abaixo das taxas de controle de utilização são processadas, enquanto aquelas acima do limite definido são rejeitadas com uma mensagem de retorno indicando que o uso da solicitação foi controlado.

Resultado desejado: Grandes picos de volume, sejam causados por aumentos repentinos de tráfego de clientes, ataques de inundação ou tempestades de novas tentativas, são reduzidos pelo controle de utilização de solicitações, permitindo que as workloads continuem com o processamento normal do volume de solicitações compatível.

Antipadrões comuns:

- Os controles de utilização de endpoint da API não são implementados ou são mantidos em valores padrão sem considerar os volumes esperados.
- Não há teste de carregamento nem limites de controle de utilização dos endpoints da API.
- Controlar a utilização de taxas de solicitações sem considerar o tamanho ou a complexidade da solicitação.
- Testar as taxas máximas de solicitação ou o tamanho máximo da solicitação, mas não testar os dois juntos.
- Os recursos não são provisionados nos mesmos limites estabelecidos nos testes.
- Os planos de uso não foram configurados nem considerados para consumidores de API de aplicação para aplicação (A2A).

- Os consumidores da fila que escalam horizontalmente não têm as configurações máximas de simultaneidade configuradas.
- A limitação de taxas por endereço IP não foi implementada.

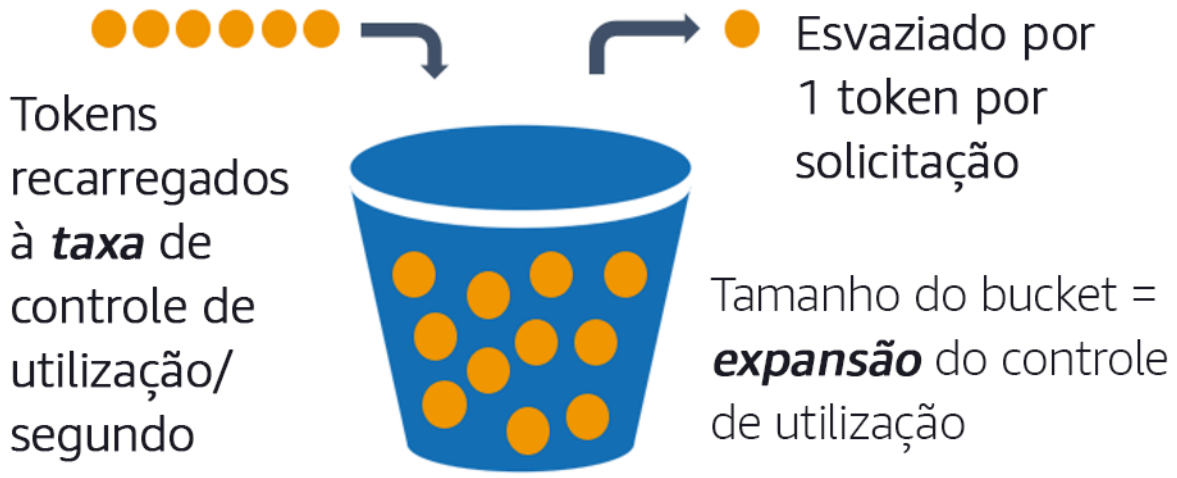
Benefícios de estabelecer esta prática recomendada: As workloads que definem limites de controle de utilização podem operar normalmente e processar a carga de solicitações aceitas com êxito em picos de volume inesperados. Os picos repentinos ou contínuos de solicitações para APIs e filas têm controle de utilização e não esgotam os recursos de processamento de solicitações. Os limites de taxas controlam a utilização de solicitantes individuais para que grandes volumes de tráfego de um único endereço IP ou consumidor de API não esgotem os recursos e afetem outros consumidores.

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

Orientação para implementação

Os serviços devem ser projetados para processar uma capacidade conhecida de solicitações; essa capacidade pode ser estabelecida por meio de testes de carga. Se as taxas de chegada de solicitações excederem os limites, a resposta apropriada sinalizará que uma solicitação teve controle de utilização. Isso permite que o consumidor resolva o erro e tente novamente mais tarde.

Quando seu serviço exigir uma implementação de controle de utilização, considere implementar o algoritmo de bucket de token, em que um token é contabilizado para uma solicitação. Os tokens são recarregados a uma taxa de controle de utilização por segundo e esvaziados de forma assíncrona por meio de um token por solicitação.



O algoritmo do bucket de token.

O Amazon API Gateway implementa o algoritmo do bucket de token de acordo com os limites da conta e da região e pode ser configurado por cliente com planos de uso. Além disso, o Amazon Simple Queue Service (Amazon SQS) e o Amazon Kinesis podem armazenar solicitações em buffer para suavizar a taxa de solicitações e permitir taxas de controle de utilização mais altas para solicitações que podem ser atendidas. Por fim, você pode implementar a limitação de taxa com o AWS WAF para controlar a utilização de consumidores de API específicos que geram uma carga excepcionalmente alta.

Etapas da implementação

É possível configurar o API Gateway com limites de controle de utilização para suas APIs e retornar erros "`429 Muitas solicitações`" quando os limites são excedidos. Você pode usar o AWS WAF com seus endpoints do AWS AppSync e do API Gateway para habilitar o limite de taxa por endereço IP. Além disso, se seu sistema tolerar o processamento assíncrono, será possível colocar mensagens em uma fila ou em um fluxo para acelerar as respostas aos clientes do serviço, o que permite que você atinja taxas de controle de utilização mais altas.

Com o processamento assíncrono, ao configurar o Amazon SQS como fonte de eventos para o AWS Lambda, é possível configurar a simultaneidade máxima para evitar que altas taxas de eventos consumam a cota de execução simultânea da conta disponível necessária para outros serviços em sua workload ou conta.

Embora o API Gateway ofereça uma implementação gerenciada do bucket de token, em casos em que não é possível usar o API Gateway, você pode utilizar as implementações de código aberto específicas da linguagem (veja exemplos relacionados em Recursos) do bucket de token para seus serviços.

- Entenda e configure limites de controle de utilização do API Gateway em nível de conta por região, API por estágio e chave de API por nível do plano de uso.
- Aplique regras de controle de utilização de taxas do AWS WAF para endpoints do API Gateway e do AWS AppSync a fim de se proteger contra inundações e bloquear IPs mal-intencionados. As regras de controle de utilização de taxas também podem ser configuradas em chaves de API do AWS AppSync para consumidores A2A.
- Decida se você precisa de mais controle de limitação do que limitação de taxas para APIs do AWS AppSync e, em caso afirmativo, configure um API Gateway na frente do seu endpoint do AWS AppSync.
- Quando filas do Amazon SQS são configuradas como gatilhos para os consumidores da fila do Lambda, defina a simultaneidade máxima como um valor que processe o suficiente para atender

aos seus objetivos de nível de serviço, mas não consuma limites de simultaneidade que afetem outras funções do Lambda. Considere definir a simultaneidade reservada em outras funções do Lambda na mesma conta e região ao consumir filas com o Lambda.

- Use o API Gateway com integrações de serviços nativos para Amazon SQS ou Kinesis para armazenar solicitações em buffer.
- Se você não puder usar o API Gateway, consulte bibliotecas específicas de linguagens para implementar o algoritmo do bucket de token para sua workload. Confira a seção de exemplos e faça sua própria pesquisa para encontrar uma biblioteca adequada.
- Teste os limites que você planeja definir ou permitir que sejam aumentados e documente os limites testados.
- Não aumente os limites além do que você estabeleceu nos testes. Ao aumentar um limite, verifique se os recursos provisionados já são equivalentes ou maiores do que os dos cenários de teste antes de aplicar o aumento.

Recursos

Práticas recomendadas relacionadas:

- REL04-BP03 Fazer um trabalho constante
- REL05-BP03 Controlar e limitar as chamadas de repetição

Documentos relacionados:

- Amazon API Gateway: controlar a utilização das solicitações de API para um melhor throughput
- AWS WAF: instrução de regra baseada em taxas
- Introdução da máxima simultaneidade do AWS Lambda ao usar o Amazon SQS como fonte de eventos
- AWS Lambda: simultaneidade máxima

Exemplos relacionados:

- As três regras mais importantes baseadas em taxas do AWS WAF
- Java Bucket4j
- Bucket de tokens do Python
- Bucket de tokens do Node

- Limitação da taxa de segmentação do sistema .NET

Vídeos relacionados:

- Implementing GraphQL API security best practices with AWS AppSync (Implementação das práticas recomendadas de segurança da API GraphQL com AWS AppSync)

Ferramentas relacionadas:

- O Amazon API Gateway
- AWS AppSync
- Amazon SQS
- Amazon Kinesis
- AWS WAF

REL05-BP03 Controlar e limitar as chamadas de repetição

Use o recuo exponencial para repetir as solicitações em intervalos progressivamente maiores entre cada nova repetição. Introduza o jitter entre as repetições para tornar os intervalos de repetição aleatórios. Limite o número máximo de repetições.

Resultado desejado: Os componentes típicos em um sistema de software distribuído incluem servidores, load balancers, bancos de dados e servidores DNS. Durante a operação normal, esses componentes podem responder a solicitações com erros temporários ou limitados, além de erros que seriam persistentes, independentemente de repetições. Quando os clientes fazem solicitações aos serviços, elas consomem recursos, incluindo memória, threads, conexões, portas ou quaisquer outros recursos limitados. Controlar e limitar as repetições é uma estratégia para liberar e minimizar o consumo de recursos para que os componentes do sistema sob pressão não fiquem sobrecarregados.

Quando as solicitações do cliente atingem o tempo limite ou recebem respostas de erro, ele deve determinar se deve ou não tentar novamente. Se tentar novamente, ele o fará com um recuo exponencial com jitter e um valor máximo de repetição. Como resultado, os serviços e os processos de back-end recebem alívio da carga e do tempo de recuperação automática, ocasionando uma recuperação mais rápida e atendimento bem-sucedido das solicitações.

Antipadrões comuns:

- Implementar repetições sem adicionar recuo exponencial, jitter e valores máximos de repetição. O recuo e o jitter ajudam a evitar picos artificiais de tráfego devido a repetições coordenadas involuntariamente em intervalos comuns.
- Implementar repetições sem testar seus efeitos ou presumir que repetições já estejam incorporadas a um SDK sem testar cenários de repetição.
- Não entender os códigos de erro publicados das dependências, ocasionando a repetição de todos os erros, inclusive aqueles com uma causa clara que indica falta de permissão, erro de configuração ou outra condição que, previsivelmente, não será resolvida sem intervenção manual.
- Não abordar práticas de observabilidade, incluindo monitoramento e alertas sobre falhas repetidas de serviço para que os problemas subjacentes sejam divulgados e possam ser resolvidos.
- Desenvolver mecanismos de repetição personalizados quando os recursos de repetição integrados ou de terceiros são suficientes.
- Tentar novamente em várias camadas da pilha de aplicações de uma forma que agrava as tentativas de repetição, consumindo ainda mais recursos em uma tempestade de repetições. Entenda como esses erros afetam sua aplicação, as dependências nas quais você confia e implemente repetições em apenas um nível.
- Repetir chamadas de serviço que não são idempotentes, causando efeitos colaterais inesperados, como resultados duplicados.

Benefícios de estabelecer esta prática recomendada: As repetições ajudam os clientes a obter os resultados desejados quando as solicitações falham, mas também consomem mais tempo do servidor para obter as respostas bem-sucedidas que eles desejam. Quando as falhas são raras ou transitórias, as repetições funcionam bem. Quando as falhas são causadas pela sobrecarga de recursos, as repetições podem piorar as coisas. Adicionar um recuo exponencial com jitter às repetições do cliente permite que os servidores se recuperem quando as falhas são causadas pela sobrecarga de recursos. O jitter evita o alinhamento das solicitações em picos, e o recuo diminui o escalonamento de carga causado pela adição de repetições à carga normal da solicitação. Por fim, é importante configurar um número máximo de repetições ou o tempo decorrido para evitar a criação de backlogs que produzam falhas metaestáveis.

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

Orientação para implementação

Controle e limite as chamadas de repetição. Use o recuo exponencial para tentar novamente após intervalos progressivamente mais longos. Introduza jitter para tornar esses intervalos de repetição aleatórios e limite o número máximo de repetições.

Alguns AWS SDKs implementam repetições e recuo exponencial por padrão. Use essas implementações integradas da AWS quando aplicável em sua workload. Implemente uma lógica semelhante em sua workload ao chamar serviços que sejam idempotentes e em que repetições melhorem a disponibilidade do cliente. Decida quais são os tempos limite e quando parar de tentar novamente com base no seu caso de uso. Crie e exercite cenários de teste para esses casos de uso de repetições.

Etapas da implementação

- Determine a camada ideal em sua pilha de aplicações para implementar repetições para os serviços dos quais sua aplicação depende.
- Conheça os SDKs existentes que implementam estratégias comprovadas de repetição com retrocesso exponencial e jitter para a linguagem de sua escolha e dê preferência a esses SDKs em vez de escrever suas próprias implementações de repetição.
- Verifique se os serviços são idempotentes antes de implementar repetições. Depois que as repetições forem implementadas, elas devem ser testadas e exercitadas regularmente na produção.
- Ao chamar APIs de serviço da AWS, use os AWS SDKs e o AWS CLI e entenda as opções de configuração de repetições. Determine se os padrões funcionam para seu caso de uso, teste e ajuste conforme necessário.

Recursos

Práticas recomendadas relacionadas:

- REL04-BP04 Fazer com que todas as respostas sejam idempotentes
- REL05-BP02 Controlar a utilização de solicitações
- REL05-BP04 Antecipar-se à falha e filas limitadas
- REL05-BP05 Definir tempos limite do cliente
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

Documentos relacionados:

- Repetições de erros e recuo exponencial na AWS
- A Amazon Builders' Library: tempos limite, novas tentativas e recuo com tremulação
- Recuo exponencial e jitter
- Tornar as tentativas seguras com APIs idempotentes

Exemplos relacionados:

- Repetição Spring
- Repetição Resilience4j

Vídeos relacionados:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328) (Repetição, recuo e jitter: AWS re:Invent 2019: Introdução à biblioteca de criadores da Amazon (DOP328)

Ferramentas relacionadas:

- AWS SDKs e ferramentas: comportamento de repetição
- AWS Command Line Interface: repetições da AWS CLI

REL05-BP04 Antecipar-se à falha e filas limitadas

Quando um serviço não consegue responder com êxito a uma solicitação, antecipe-se à falha. Isso permite a liberação dos recursos associados a uma solicitação e possibilita que o serviço se recupere se estiver ficando sem recursos. Antecipar-se à falha é um padrão de design de software bem estabelecido que pode ser utilizado para criar workloads altamente confiáveis na nuvem. As filas também correspondem a um padrão de integração empresarial bem estabelecido que pode facilitar o carregamento e permitir que os clientes liberem recursos quando o processamento assíncrono pode ser tolerado. Quando um serviço consegue responder com êxito em condições normais, mas falha quando a taxa de solicitações é muito alta, use uma fila para armazenar solicitações em buffer. No entanto, não permita a formação de backlogs de filas longas que possam ocasionar o processamento de solicitações antigas das quais um cliente já desistiu.

Resultado desejado: Quando os sistemas enfrentam contenção de recursos, tempos limite, exceções ou falhas de causa desconhecida que tornam os objetivos de nível de serviço inatingíveis, as estratégias de antecipação a falhas permitem uma recuperação mais rápida do sistema. Sistemas que precisam absorver picos de tráfego e acomodar o processamento assíncrono podem melhorar a confiabilidade ao permitir que os clientes liberem solicitações rapidamente usando filas para armazenar solicitações em buffer para serviços de back-end. Ao armazenar solicitações em filas, estratégias de gerenciamento de filas são implementadas para evitar backlogs intransponíveis.

Antipadrões comuns:

- Implementar filas de mensagens, mas não configurar filas de mensagens não entregues (DLQ) ou alarmes em volumes DLQ para detectar quando um sistema está em falha.
- Não medir a idade das mensagens em uma fila, uma medida de latência para entender quando os consumidores da fila estão ficando para trás ou cometendo erros, ocasionando repetições.
- Não limpar mensagens pendentes de uma fila, quando não há utilidade em processar essas mensagens se a necessidade empresarial deixar de existir.
- Configurar filas do tipo “first in first out” (FIFO) quando filas do tipo “last in first out” (LIFO) atenderia melhor às necessidades do cliente, por exemplo, quando a ordenação rigorosa não é necessária e o processamento de backlog está atrasando todas as solicitações novas e urgentes, ocasionando violação dos níveis de serviço de todos os clientes.
- Expor filas internas aos clientes em vez de expor APIs que gerenciem a entrada de trabalho e coloquem as solicitações em filas internas.
- Combinar muitos tipos de solicitações de trabalho em uma única fila, o que pode agravar as condições de backlog ao distribuir a demanda de recursos entre os tipos de solicitação.
- Processar solicitações complexas e simples na mesma fila, apesar da necessidade de monitoramento, tempos limite e alocação de recursos diferentes.
- Não validar entradas ou usar afirmações para implementar mecanismos de antecipação à falha em software que agreguem exceções a componentes de nível superior que podem lidar com erros sem problemas.
- Não remover recursos com defeito do roteamento de solicitações, principalmente quando as falhas estão emitindo êxitos e falhas em decorrência de travamento e reinicialização, falha de dependência intermitente, capacidade reduzida ou perda de pacotes de rede.

Benefícios de estabelecer esta prática recomendada: Sistemas que se antecipam à falha são mais fáceis de depurar e corrigir e geralmente expõem problemas de codificação e configuração antes que

as versões sejam publicadas em produção. Os sistemas que incorporam estratégias eficazes de filas oferecem maior resiliência e confiabilidade a picos de tráfego e às condições intermitentes de falha do sistema.

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

Orientação para implementação

As estratégias de antecipação à falha podem ser codificadas em soluções de software e configuradas em infraestrutura. Além de se anteciparem à falha, as filas são uma técnica arquitetônica simples, mas poderosa, para dissociar os componentes do sistema e facilitar o carregamento. O Amazon CloudWatch oferece recursos para monitorar e alertar sobre falhas. Quando se sabe que um sistema está falhando, estratégias de mitigação podem ser invocadas, inclusive evitar recursos afetados. Quando os sistemas implementam filas com o Amazon SQS e outras tecnologias de fila para facilitar o carregamento, eles devem considerar como gerenciar os backlogs de filas, bem como as falhas no consumo de mensagens.

Etapas da implementação

- Implemente afirmações programáticas ou métricas específicas em seu software e use-as para alertar explicitamente sobre problemas do sistema. O Amazon CloudWatch ajuda você a criar métricas e alarmes com base no padrão de log da aplicação e na instrumentação do SDK.
- Use métricas e alarmes do CloudWatch para eliminar recursos danificados que estão aumentando a latência no processamento ou falhando repetidamente no processamento das solicitações.
- Use o processamento assíncrono criando APIs para aceitar e anexar solicitações às filas internas usando o Amazon SQS e, depois, responder ao cliente que produz a mensagem com uma mensagem de êxito para que o cliente possa liberar recursos e prosseguir com outros trabalhos enquanto os consumidores da fila de back-end processam as solicitações.
- Avalie e monitore a latência do processamento da fila produzindo uma métrica do CloudWatch sempre que retirar uma mensagem de uma fila, comparando o momento presente com o carimbo de data/hora da mensagem.
- Quando falhas impedem o processamento bem-sucedido de mensagens ou geram picos de tráfego em volumes que não podem ser processados de acordo com acordos de serviço, deixe de lado o tráfego antigo ou excedente para uma fila de transbordamento. Isso permite o processamento prioritário de trabalhos novos e antigos, quando há capacidade disponível. Essa técnica é uma aproximação do processamento LIFO e permite o processamento normal do sistema para todos os novos trabalhos.

- Use filas de mensagens não entregues ou de redirecionamento para mover mensagens que não podem ser processadas do backlog para um local que possa ser pesquisado e resolvido posteriormente.
- Tente novamente ou, quando possível, elimine as mensagens antigas comparando o momento presente com o carimbo de data/hora da mensagem e descartando as mensagens que não são mais relevantes para o cliente solicitante.

Recursos

Práticas recomendadas relacionadas:

- REL04-BP02 Implementar dependências com acoplamento fraco
- REL05-BP02 Controlar a utilização de solicitações
- REL05-BP03 Controlar e limitar as chamadas de repetição
- REL06-BP02 Definir e calcular as métricas (agregação)
- REL06-BP07 Monitorar o rastreamento completo das solicitações por meio de seu sistema

Documentos relacionados:

- Evitar backlogs de fila intransponíveis
- Falha rápida
- Como evitar um aumento do backlog de mensagens na minha fila do Amazon SQS?
- Elastic Load Balancing: mudança de zona
- Controlador de recuperação de aplicações Amazon Route 53: controle de roteamento para failover de tráfego

Exemplos relacionados:

- Padrões de integração empresarial: canal de mensagens não entregues

Vídeos relacionados:

- AWS re:Invent 2022 - Operating highly available Multi-AZ applications (re:Invent 2022: operar aplicações Multi-AZ altamente disponíveis)

Ferramentas relacionadas:

- Amazon SQS
- Amazon MQ
- AWS IoT Core
- O Amazon CloudWatch

REL05-BP05 Definir tempos limite do cliente

Defina tempos limite adequados para conexões e solicitações, verifique-os sistematicamente e não confie nos valores padrão, pois eles não estão cientes das especificações da workload.

Resultado desejado: Os tempos limite do cliente devem considerar o custo para o cliente, o servidor e a workload associados à espera por solicitações que levam um tempo anormal para serem concluídas. Como não é possível saber a causa exata de nenhum tempo limite, os clientes devem usar o conhecimento dos serviços para desenvolver expectativas de causas prováveis e prazos apropriados.

As conexões do cliente atingem o tempo limite com base nos valores configurados. Depois de encontrar um tempo limite, os clientes tomam a decisão de recuar e tentar novamente ou abrir um disjuntor. Esses padrões evitam a emissão de solicitações que podem exacerbar uma condição de erro subjacente.

Antipadrões comuns:

- Não estar ciente dos tempos limite do sistema ou dos tempos limite padrão.
- Não estar ciente do tempo normal de conclusão da solicitação.
- Não estar ciente das possíveis causas das solicitações levarem muito tempo para serem concluídas ou dos custos de performance do cliente, do serviço ou da workload associados à espera por essas conclusões.
- Não estar ciente da probabilidade de uma rede danificada fazer com que uma solicitação falhe somente quando o tempo limite é atingido e dos custos para a performance do cliente e da workload por não adotar um tempo limite mais curto.
- Não testar cenários de tempo limite tanto para conexões quanto para solicitações.
- Definir tempos limite muito altos, o que pode resultar em longos tempos de espera e aumentar a utilização de recursos.
- Definir tempos limite muito baixos, gerando falhas artificiais.

- Ignorar padrões para lidar com erros de tempo limite para chamadas remotas, como disjuntores e novas tentativas.
- Não considerar o monitoramento de taxas de erro de chamadas de serviço, objetivos de nível de serviço para latência e valores atípicos de latência. Essas métricas podem fornecer informações sobre tempos limite agressivos ou permissivos.

Benefícios de estabelecer esta prática recomendada: Os tempos limite de chamadas remotas são configurados e os sistemas são projetados para lidar com os tempos limite normalmente, de forma que os recursos sejam conservados quando as chamadas remotas respondem de forma anormalmente lenta e os erros de tempo limite são tratados normalmente pelos clientes do serviço.

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

Orientação para implementação

Defina um tempo limite de conexão e um tempo limite de solicitação em qualquer chamada de dependência de serviço e, geralmente, em qualquer chamada entre processos. Muitas frameworks oferecem recursos de tempo limite integrados, mas tenha cuidado, pois algumas têm valores padrão infinitos ou superiores ao aceitável para seus objetivos de serviço. 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.

Considere o seguinte ao determinar as estratégias de tempo limite:

- As solicitações podem levar mais tempo do que o normal para serem processadas devido ao conteúdo, a deficiências em um serviço de destino ou a uma falha na partição de rede.
- Solicitações com conteúdo anormalmente caro podem consumir recursos desnecessários do servidor e do cliente. Nesse caso, reduzir o tempo limite dessas solicitações e não tentar novamente pode preservar os recursos. Os serviços também devem se proteger de conteúdo anormalmente caro com limitações e tempos limite do servidor.
- Solicitações que demoram muito devido a uma falha no serviço podem expirar e ser repetidas. Deve-se considerar os custos do serviço para a solicitação e a nova tentativa, mas se a causa for uma deficiência localizada, uma nova tentativa provavelmente não será cara e reduzirá o consumo de recursos do cliente. O tempo limite também pode liberar recursos do servidor, dependendo da natureza da deficiência.

- Solicitações que demoram muito para serem concluídas porque a solicitação ou a resposta não foi entregue pela rede podem expirar e ser repetidas. Como a solicitação ou a resposta não foi entregue, a falha teria sido o resultado, independentemente da duração do tempo limite. Nesse caso, o tempo limite não liberará recursos do servidor, mas liberará recursos do cliente e melhorará a performance da workload.

Aproveite os padrões de design bem estabelecidos, como novas tentativas e disjuntores, para lidar com os tempos de espera de forma eficiente e oferecer compatibilidade com abordagens de antecipação à falha. AWS SDKs e a AWS CLI permitem a configuração dos tempos limite de conexão e solicitação e as repetições com recuo exponencial e instabilidade. As funções do AWS Lambda são compatíveis com a configuração de tempos limite e com o AWS Step Functions, você pode criar disjuntores de uso de pouco código que utilizam integrações pré-incorporadas com serviços da AWS e SDKs. O AWS App Mesh Envoy oferece recursos de tempo limite e disjuntor.

Etapas da implementação

- Configure tempos limite em chamadas de serviço remoto e utilize os recursos de tempo limite de linguagem integrados ou as bibliotecas de tempo limite de código aberto.
- Quando sua workload fizer chamadas com um AWS SDK, revise a documentação para saber a configuração de tempo limite específica da linguagem.
  - Python
  - PHP
  - .NET
  - Ruby
  - Java
  - Go
  - Node.js
  - C++
- Ao usar AWS SDKs ou comandos da AWS CLI em sua workload, configure os valores de tempo limite padrão definindo os padrões de configuração da AWS de `connectTimeoutInMillis` e `tlsNegotiationTimeoutInMillis`.
- Aplique opções de linha de comando `cli-connect-timeout` e `cli-read-timeout` para controlar comandos únicos da AWS CLI para serviços da AWS.
- Monitore o tempo limite de chamadas de serviço remoto e defina alarmes para erros persistentes para que você possa lidar proativamente com cenários de erro.

- Implemente métricas do CloudWatch e detecção de anomalias do CloudWatch em taxas de erro de chamada, objetivos de nível de serviço para latência e valores atípicos de latência para fornecer informações sobre o gerenciamento de tempos limite excessivamente agressivos ou permissivos.
- Configure tempos limite em funções do Lambda.
- Os clientes do API Gateway devem implementar suas próprias repetições ao lidar com os tempos limite. O API Gateway é compatível com um tempo limite de integração de 50 milissegundos a 29 segundos para integrações posteriores e não tenta novamente quando as solicitações de integração atingem o tempo limite.
- Implemente o padrão de disjuntor para evitar fazer chamadas remotas quando o tempo limite está prestes a ser atingido. Abra o circuito para evitar falhas nas chamadas e feche-o quando as chamadas estiverem respondendo normalmente.
- Para workloads baseadas em contêineres, analise os recursos do App Mesh Envoy para utilizar os tempos limite e os disjuntores integrados.
- Use o AWS Step Functions para criar disjuntores de pouco uso de código para chamadas de serviço remoto, especialmente ao chamar AWS SDKs nativos e integrações do Step Functions compatíveis para simplificar sua workload.

Recursos

Práticas recomendadas relacionadas:

- REL05-BP03 Controlar e limitar as chamadas de repetição
- REL05-BP04 Antecipar-se à falha e filas limitadas
- REL06-BP07 Monitorar o rastreamento completo das solicitações por meio de seu sistema

Documentos relacionados:

- AWS SDK: repetições e tempos limite
- A Amazon Builders' Library: tempos limite, novas tentativas e recuo com tremulação
- Cotas do Amazon API Gateway e notas importantes
- AWS Command Line Interface: opções de linha de comando
- AWS SDK for Java 2.x: configurar tempos limite de API
- AWS Botocore usando o objeto de configuração e a referência de configuração
- AWS SDK for .NET: repetições e tempos limite

- AWS Lambda: configurar as opções de função do Lambda

Exemplos relacionados:

- Usar o padrão do disjuntor com o AWS Step Functions e o Amazon DynamoDB
- Martin Fowler: CircuitBreaker

Ferramentas relacionadas:

- AWS SDKs
- As funções do AWS Lambda
- Amazon SQS
- AWS Step Functions
- AWS Command Line Interface

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.

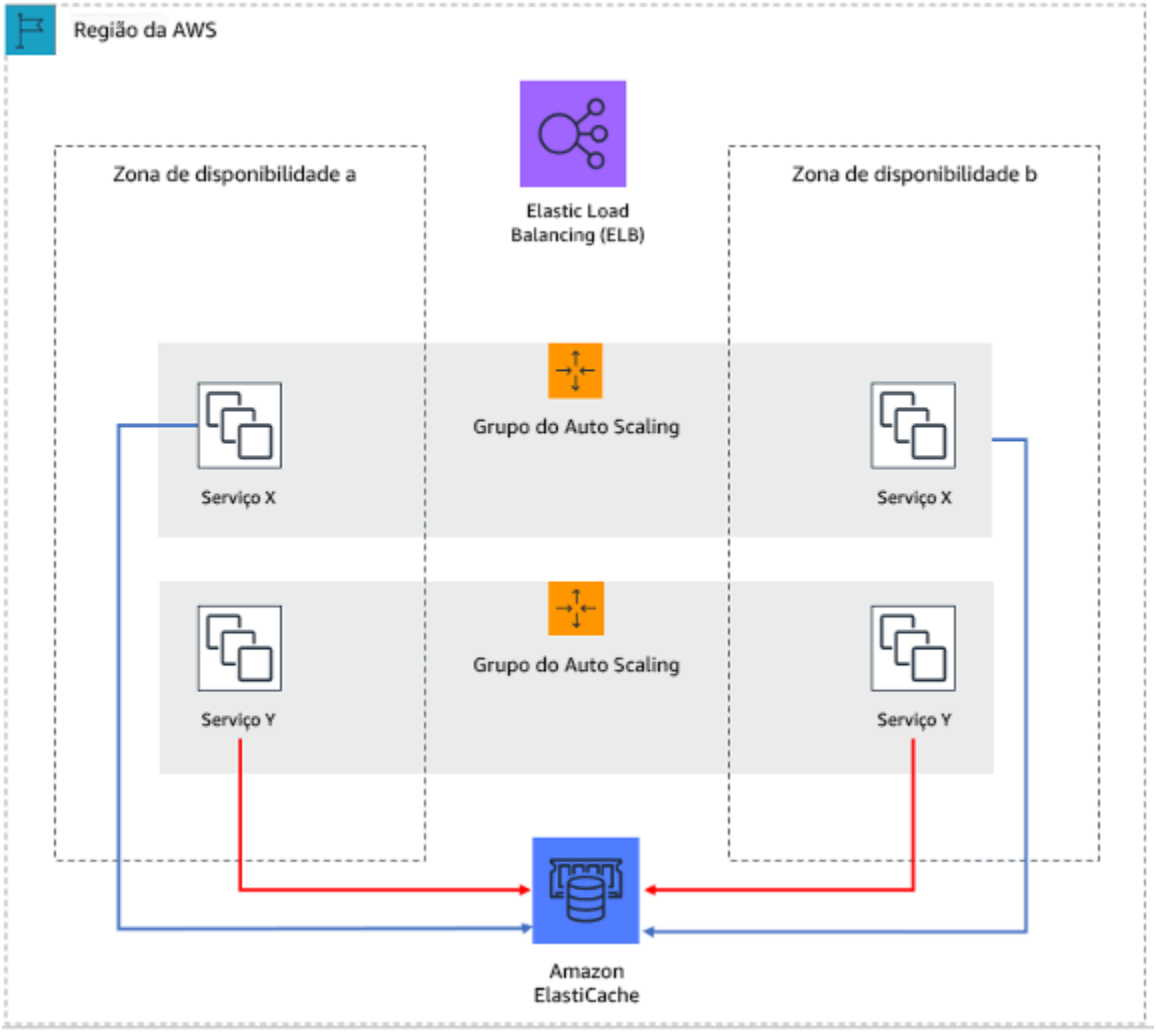


Figura 7: Nesta aplicação Web sem estado, o estado da sessão é descarregado para o Amazon ElastiCache.

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.

As medidas emergenciais funcionam com a desativação, o controle de utilização ou a alteração do comportamento dos componentes ou das dependências com o uso de mecanismos conhecidos e testados. Isso pode aliviar as deficiências da workload decorrentes da exaustão dos recursos

provocada por aumentos inesperados na demanda e reduzir o impacto de falhas em componentes não essenciais da workload.

Resultado desejado: ao implementar medidas de emergência, é possível estabelecer processos bem conhecidos para manter a disponibilidade dos componentes essenciais na workload. A workload deve se degradar normalmente e continuar desempenhando suas funções essenciais aos negócios durante a ativação de uma medida emergencial. Para obter mais detalhes sobre a degradação simples, consulte REL05-BP01 Implementar uma degradação simples para transformar dependências rígidas aplicáveis em dependências flexíveis.

Antipadrões comuns:

- A falha de dependências não essenciais afeta a disponibilidade da workload principal.
- Não testar ou verificar o comportamento dos componentes essenciais durante a deterioração de componentes não essenciais.
- Não há critérios claros e determinísticos definidos para ativação ou desativação de uma medida emergencial.

Benefícios do estabelecimento desta prática recomendada: a implementação de medidas emergenciais pode melhorar a disponibilidade dos componentes essenciais na workload fornecendo aos resolvedores processos estabelecidos para responder a picos inesperados na demanda ou a falhas de dependências não essenciais.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

- Identifique os componentes essenciais na workload.
- Projete e arquitete os componentes essenciais na workload para resistirem à falha de componentes não essenciais.
- Conduza testes para validar o comportamento dos componentes essenciais durante a falha de componentes não essenciais.
- Defina e monitore métricas ou acionadores relevantes para iniciar procedimentos de medida emergencial.
- Defina os procedimentos (manuais ou automatizados) que compõem a medida emergencial.

Etapas da implementação

- Identificar os componentes essenciais aos negócios na workload.
  - Cada componente técnico na workload deve ser mapeado para a função de negócios relevante e classificado como essencial ou não essencial. Para obter exemplos de funcionalidades essenciais e não essenciais na Amazon, consulte Any Day Can Be Prime Day: How Amazon.com Search Uses Chaos Engineering to Handle Over 84K Requests Per Second.
  - Essa é uma decisão técnica e de negócios e varia de acordo com a organização e a workload.
- Projete e arquitete os componentes essenciais na workload para resistirem à falha de componentes não essenciais.
  - Durante a análise de dependências, considere todos os possíveis modos de falha e verifique se os mecanismos de medida emergencial fornecem a funcionalidade essencial aos componentes subsequentes.
- Conduza testes para validar o comportamento dos componentes essenciais durante a ativação das medidas emergenciais.
  - Evite o comportamento bimodal. Para obter mais detalhes, consulte REL11-BP05 Usar estabilidade estática para evitar o comportamento bimodal.
- Defina, monitore e emita alertas sobre as métricas relevantes para iniciar o procedimento de medida emergencial.
  - A descoberta das métricas certas a serem monitoradas depende da workload. Alguns exemplos de métricas são a latência ou o número de solicitações com falha feitas para uma dependência.
- Defina os procedimentos, manuais ou automatizados, que compõem a medida emergencial.
  - Isso pode incluir mecanismos como descarte de carga, controle de utilização de solicitações ou implementação de degradação simples.

Recursos

Práticas recomendadas relacionadas:

- REL05-BP01 Implementar uma degradação simples para transformar dependências rígidas aplicáveis em dependências flexíveis
- REL05-BP02 Controlar a utilização de solicitações
- REL11-BP05 Usar estabilidade estática para evitar o comportamento bimodal

Documentos relacionados:

- Automatizar uma implantação prática e sem intervenção manual
- Any Day Can Be Prime Day: How Amazon.com Search Uses Chaos Engineering to Handle Over 84K Requests Per Second

Vídeos relacionados:

- AWS re:Invent 2020: Reliability, consistency, and confidence through immutability

Gerenciamento de alterações

Perguntas

- CONFIABILIDADE 6. Como monitorar recursos de workload?
- CONFIABILIDADE 7. Como projetar sua workload para se adaptar às mudanças na demanda?
- CONFIABILIDADE 8. Como implementar uma alteração?

CONFIABILIDADE 6. Como monitorar recursos de workload?

Os logs e as métricas são uma ferramenta poderosa para saber a integridade de sua workload. Você pode configurar sua workload para monitorar logs e métricas e enviar notificações quando os limites forem ultrapassados ou em caso de eventos importantes. O monitoramento permite que sua workload 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 Envie 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 de 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 Envie notificações (processamento e emissão de alarmes em tempo real)

Quando as organizações detectam possíveis problemas, elas enviam notificações e alertas em tempo real para a equipe e os sistemas apropriados para responder de forma rápida e eficaz a esses problemas.

Resultado desejado: respostas rápidas a eventos operacionais são possíveis por meio da configuração de alarmes relevantes com base nas métricas de serviços e aplicações. Quando os limites do alarme são violados, o pessoal e os sistemas apropriados são notificados para que possam resolver os problemas subjacentes.

Antipadrões comuns:

- Configuração de alarmes com um limite excessivamente alto, resultando em falha no envio de notificações vitais.
- Configurar alarmes com um limite muito baixo, ocasionando inatividade diante de alertas importantes devido ao ruído de notificações excessivas.
- Não atualizar os alarmes e seu limite quando o uso muda.
- Para alarmes mais bem abordados por meio de ações automatizadas, enviar a notificação ao pessoal em vez de gerar a ação automatizada gera o envio excessivo de notificações.

Benefícios de estabelecer esta prática recomendada: O envio de notificações e alertas em tempo real para o pessoal e os sistemas apropriados permite a detecção precoce de problemas e respostas rápidas aos incidentes operacionais.

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

Orientação para implementação

As workloads devem ser equipadas com processamento e alarmes em tempo real para melhorar a detectabilidade de problemas que possam afetar a disponibilidade da aplicação e servir como gatilhos para respostas automatizadas. As organizações podem realizar processamento e alarmes em tempo real criando alertas com métricas definidas para receber notificações sempre que eventos significativos ocorrerem ou quando uma métrica ultrapassar um limite.

Amazon CloudWatch permite criar alarmes de métricas e compostos usando alarmes do CloudWatch com base em limite estático, detecção de anomalias e outros critérios. Para obter mais detalhes sobre os tipos de alarme que você pode configurar usando o CloudWatch, consulte a seção de alarmes da documentação do CloudWatch.

É possível criar visualizações personalizadas de métricas e alertas dos recursos da AWS para as equipes usando painéis do CloudWatch. As páginas iniciais personalizáveis no console do CloudWatch permitem que você monitore seus recursos em uma única visualização em várias regiões.

Os alarmes podem realizar uma ou mais ações, como enviar uma notificação a um tópico do Amazon SNS, realizando uma ação do Amazon EC2 ou do Amazon EC2 Auto Scaling , ou criando um OpsItem ou incidente no AWS Systems Manager.

O Amazon CloudWatch usa o Amazon SNS para enviar notificações quando o alarme muda de estado, fornecendo a entrega de mensagens dos publicadores (produtores) para os assinantes (consumidores). Para obter mais detalhes sobre como configurar notificações do Amazon SNS, consulte Configuring Amazon SNS.

O CloudWatch envia eventos do EventBridge segurança sempre que um alarme do CloudWatch é criado, atualizado, excluído ou o estado muda. É possível usar o EventBridge com esses eventos para criar regras que realizam ações, como enviar uma notificação sempre que o estado de um alarme mudar ou acionar eventos automaticamente na conta usando o Systems Manager Automation.

Quando você deve usar o EventBridge ou o Amazon SNS?

Tanto o EventBridge quanto o Amazon SNS podem ser usados para desenvolver aplicações orientadas a eventos, e sua escolha dependerá de suas necessidades específicas.

O Amazon EventBridge é recomendado quando você deseja criar uma aplicação que reaja a eventos de suas próprias aplicações, aplicações SaaS e serviços da AWS. O EventBridge é o único serviço baseado em eventos que se integra diretamente com parceiros SaaS de terceiros. O EventBridge também ingere automaticamente eventos de mais de 200 serviços da AWS sem exigir que os desenvolvedores criem recursos em suas contas.

O EventBridge usa uma estrutura definida baseada em JSON para eventos e ajuda você a criar regras que são aplicadas em todo o corpo do evento para selecionar eventos a serem encaminhados a um destino. O EventBridge no momento é compatível com mais de vinte serviços da AWS como destinos, incluindo AWS Lambda, o Amazon SQS, Amazon SNS, Amazon Kinesis Data Streamse o Amazon Data Firehose.

O Amazon SNS é recomendado para aplicações que precisam de alta distribuição (milhares ou milhões de endpoints). Um padrão comum que vemos é que os clientes usam o Amazon SNS como destino para a regra a fim de filtrar os eventos de que precisam e distribuí-los para vários endpoints.

As mensagens não são estruturadas e podem estar em qualquer formato. O Amazon SNS permite o encaminhamento de mensagens a seis tipos diferentes de destinos, incluindo Lambda, Amazon SQS, endpoints HTTP/S, SMS, push de dispositivos móveis e e-mail. A latência típica do Amazon SNS é inferior a 30 milissegundos. Uma ampla variedade de serviços da AWS envia ao Amazon

SNS mensagens configurando o serviço para fazer isso (mais de trinta, incluindo o Amazon EC2, o Amazon S3e o Amazon RDS).

Etapas da implementação

1. Crie um alarme usando alarmes do Amazon CloudWatch.
    a. Um alarme de métrica monitora uma única métrica do CloudWatch ou uma expressão dependente de métricas do CloudWatch. O alarme inicia uma ou mais ações com base no valor da métrica ou expressão em comparação com um limite em vários intervalos de tempo. A ação pode consistir em enviar uma notificação a um tópico do Amazon SNS, realizando uma ação do Amazon EC2 ou do Amazon EC2 Auto Scaling , ou criando um OpsItem ou incidente no AWS Systems Manager.
    b. Um alarme composto consiste em uma expressão de regra que considera as condições de alarme de outros alarmes que você criou. O alarme composto só entra no estado de alarme se todas as condições da regra forem atendidas. Os alarmes especificados na expressão da regra de um alarme composto podem incluir alarmes de métricas e outros alarmes compostos. Alarmes compostos podem enviar notificações do Amazon SNS quando o estado muda e podem criar Systems Manager OpsItems ou incidentes quando entram no estado de alarme, mas não conseguem realizar ações do Amazon EC2 ou do Auto Scaling.
2. Configure o Notificações do Amazon SNS. Ao criar um alarme do CloudWatch, é possível incluir um tópico do Amazon SNS para enviar uma notificação quando o alarme mudar de estado.
3. Crie regras no EventBridge que corresponde aos alarmes do CloudWatch especificados. Cada regra é compatível com vários destinos, incluindo funções do Lambda. Por exemplo, você pode definir um alarme que é iniciado quando o espaço disponível em disco está acabando, o que aciona uma função do Lambda por meio de uma regra do EventBridge para limpar o espaço. Para obter mais detalhes sobre destinos do EventBridge, consulte EventBridge targets.

Recursos

Práticas recomendadas relacionadas ao Well-Architected:

- REL06-BP01 Monitorar todos os componentes da workload (geração)
- REL06-BP02 Definir e calcular as métricas (agregação)
- REL12-BP01 Usar playbooks para investigar falhas

Documentos relacionados:

- Amazon CloudWatch
- CloudWatch Logs insights
- Using Amazon CloudWatch alarms
- Using Amazon CloudWatch dashboards
- Using Amazon CloudWatch metrics (Uso de métricas do Amazon CloudWatch)
- Setting up Amazon SNS notifications
- detecção de anomalias do CloudWatch
- CloudWatch Logs data protection
- Amazon EventBridge
- Amazon Simple Notification Service

Vídeos relacionados:

- reinvent 2022 observability videos (Vídeos sobre observabilidade do AWS re:Invent 2022)
- AWS re:Invent 2022 - Observability best practices at Amazon (AWS re:Invent 2022: práticas recomendadas de observabilidade na Amazon)

Exemplos relacionados:

- Um workshop de observabilidade
- Amazon EventBridge to AWS Lambda with feedback control by Amazon CloudWatch Alarms

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.

O processamento automatizado de alarmes em tempo real é implementado para que os sistemas possam tomar medidas corretivas rapidamente e tentar evitar falhas ou degradação dos serviços quando os alarmes são acionados. As respostas automatizadas a alarmes podem incluir a substituição de componentes com falha, o ajuste da capacidade computacional, o redirecionamento do tráfego para hosts, zonas de disponibilidade ou outras regiões íntegras e a notificação dos operadores.

Resultado desejado: os alarmes em tempo real são identificados, e o processamento automatizado dos alarmes é configurado para invocar as ações apropriadas realizadas para manter os objetivos de nível de serviço e os acordos de serviço (SLAs). A automação pode variar de atividades de autorrecuperação de componentes individuais a failover de todo o site.

Antipadrões comuns:

- Não ter um inventário ou catálogo claro dos principais alarmes em tempo real.
- Não haver respostas automatizadas para alarmes essenciais (por exemplo, quando a computação está quase esgotada, ocorre o ajuste de escala automático).
- Usar ações contraditórias de resposta a alarmes.
- Não haver procedimentos operacionais padrão (SOPs) para os operadores seguirem ao receberem notificações de alerta.
- Não monitorar as alterações da configuração, pois alterações não detectadas podem causar tempo de inatividade nas workloads.
- Não haver uma estratégia para desfazer alterações não intencionais da configuração.

Benefícios do estabelecimento desta prática recomendada: a automatização do processamento de alarmes pode melhorar a resiliência do sistema. O sistema executa ações corretivas automaticamente, reduzindo as atividades manuais que permitem intervenções humanas sujeitas a erros. A workload opera, atende às metas de disponibilidade e reduz a interrupção do serviço.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

Para gerenciar alertas com eficiência e automatizar as respectivas respostas, categorize os alertas com base em sua criticidade e impacto, documente os procedimentos de resposta e planeje as respostas antes das tarefas de classificação.

Identifique tarefas que exigem ações específicas (geralmente detalhadas em runbooks) e examine todos os runbooks e manuais para determinar as tarefas que podem ser automatizadas. Geralmente, se for possível definir ações, elas poderão ser automatizadas. Se não for possível automatizar as ações, documente as etapas manuais em um SOP e treine os operadores a respeito. Conteste continuamente os processos manuais em busca de oportunidades de automação em que seja possível estabelecer e manter um plano para automatizar as respostas a alertas.

Etapas da implementação

1. Criar um inventário de alarmes: para obter uma lista de todos os alarmes, é possível utilizar a AWS CLI usando o comando do Amazon CloudWatch `describe-alarms`. Dependendo de quantos alarmes você configurou, talvez seja necessário usar paginação para recuperar um subconjunto de alarmes para cada chamada ou, alternativamente, utilizar o AWS SDK para obter os alarmes usando uma chamada de API.
2. Documentar todas as ações do alarme: atualize um runbook com todos os alarmes e as respectivas ações, independentemente de serem manuais ou automatizados. O AWS Systems Managerfornece runbooks predefinidos. Para obter mais informações sobre runbooks, consulte Working with runbooks. Para obter detalhes sobre como visualizar o conteúdo do runbook, consulte View runbook content.
3. Configurar e gerenciar ações de alarmes: para qualquer um dos alarmes que exijam uma ação, especifique a ação automatizada usando o SDK do CloudWatch. Por exemplo, é possível alterar o estado das instâncias do Amazon EC2 automaticamente com base em um alarme do CloudWatch criando e ativando ações em um alarme ou desativando ações em um alarme.

   Também é possível usar o Amazon EventBridge para responder automaticamente a eventos do sistema, como problemas de disponibilidade de aplicações ou alterações de recursos. Você pode criar regras para indicar em quais eventos tem interesse e as ações a serem executadas quando um evento corresponder a uma regra. As ações que podem ser iniciadas automaticamente incluem invocar um perfil do AWS Lambda, invocar o `Run Command` do Amazon EC2, transmitir o evento para o Amazon Kinesis Data Streams e visualizar o documento Automatizar o Amazon EC2 usando o EventBridge.
4. Procedimentos operacionais padrão (SOPs): com base nos componentes da aplicação, o AWS Resilience Hub recomenda vários modelos de SOP. É possível usar esses SOPs para documentar todos os processos que um operador deve seguir caso um alerta seja emitido. Também é possível estruturar um SOP com base nas recomendações do Resilience Hub, caso em que você precisa de uma aplicação do Resilience Hub com uma política de resiliência associada, bem como de uma avaliação histórica de resiliência em relação a essa aplicação. As recomendações para o SOP são produzidas pela avaliação de resiliência.

   O Resilience Hub trabalha com o Systems Manager para automatizar as etapas dos SOPs, fornecendo vários documentos do SSM que podem ser usados como base para esses SOPs. Por exemplo, o Resilience Hub pode recomendar um SOP para adicionar espaço em disco com base em um documento de automação existente do SSM.

5. Realizar ações automatizadas usando o Amazon DevOps Guru: é possível usar o Amazon DevOps Guru para monitorar automaticamente recursos da aplicação em busca de comportamento anômalo e fornecer recomendações direcionadas para acelerar o tempo de identificação e de correção de problemas. Com o DevOps Guru, é possível monitorar fluxos de dados operacionais quase em tempo real de várias fontes, incluindo as métricas do Amazon CloudWatch, o AWS Config, o AWS CloudFormation e o AWS X-Ray. Também é possível usar o DevOps Guru para criar OpsItems no OpsCenter e enviar eventos ao EventBridge para automação adicional.

Recursos

Práticas recomendadas relacionadas:

- REL06-BP01 Monitorar todos os componentes da workload (geração)
- REL06-BP02 Definir e calcular as métricas (agregação)
- REL06-BP03 Envie notificações (processamento e emissão de alarmes em tempo real)
- REL08-BP01 Usar runbooks para atividades padrão, como implantação

Documentos relacionados:

- AWS Systems Manager Automation
- Creating an EventBridge Rule That Triggers on an Event from an AWS Resource
- Um workshop de observabilidade
- A Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional
- What is Amazon DevOps Guru?
- Trabalhar com documentos de automação (playbooks)

Vídeos relacionados:

- AWS re:Invent 2022: Práticas recomendadas de observabilidade na Amazon
- AWS re:Invent 2020: Automate anything with AWS Systems Manager
- Introduction to AWS Resilience Hub
- Criar sistemas de tickets personalizados para notificações do Amazon DevOps Guru
- Enable Multi-Account Insight Aggregation with Amazon DevOps Guru

Exemplos relacionados:

- Reliability Workshops (Workshops sobre confiabilidade)
- Workshop do Amazon CloudWatch e do Systems Manager

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 de seu sistema

Rastreie as solicitações à medida que elas são processadas por meio de componentes de serviço para que as equipes de produto possam analisar e depurar problemas com maior facilidade e melhorar a performance.

Resultado desejado: As workloads com rastreamento abrangente em todos os componentes são fáceis de depurar, o que melhora o tempo médio até a resolução (MTTR) de erros e latência simplificando a descoberta da causa raiz. O rastreamento completo reduz o tempo necessário para descobrir os componentes afetados e detalhar as causas raiz dos erros ou da latência.

Antipadrões comuns:

- O rastreamento é usado para alguns componentes, mas não para todos. Por exemplo, sem rastrear o AWS Lambda, as equipes podem não entender claramente a latência causada por partidas a frio em uma workload com picos.
- Canários sintéticos ou monitoramento de usuário real (RUM) não são configurados com rastreamento. Sem canários ou RUM, a telemetria de interação com o cliente é omitida da análise de rastreamento, gerando um perfil de performance incompleto.
- As workloads híbridas incluem ferramentas de rastreamento nativas da nuvem e de terceiros, mas ainda não foram tomadas medidas eletivas e integram totalmente uma única solução de rastreamento. Com base na solução de rastreamento escolhida, os SDKs de rastreamento nativos de nuvem devem ser usados para instrumentar componentes que não são nativos de nuvem ou ferramentas de terceiros devem ser configuradas para ingerir a telemetria de rastreamento nativa de nuvem.

Benefícios de estabelecer esta prática recomendada: Quando as equipes de desenvolvimento são alertadas sobre problemas, elas podem ter uma visão completa das interações dos componentes do sistema, incluindo a correlação componente por componente com registros em log, performance e falhas. Como o rastreamento facilita a identificação visual das causas raiz, menos tempo é gasto na investigação delas. As equipes que entendem detalhadamente as interações dos componentes tomam decisões melhores e mais rápidas ao resolver problemas. Decisões como quando invocar o failover de recuperação de desastres (DR) ou onde melhor implementar estratégias de autorrecuperação podem ser aprimoradas com a análise de rastreamentos de sistemas e aumentar a satisfação do cliente com seus serviços.

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

Orientação para implementação

As equipes que operam aplicações distribuídas podem usar ferramentas de rastreamento para estabelecer um identificador de correlação, coletar rastreamentos de solicitações e criar mapas de serviço dos componentes conectados. Todos os componentes da aplicação devem ser incluídos nos rastreamentos de solicitações, incluindo clientes de serviços, gateways de middleware e barramentos de eventos, componentes computacionais e armazenamento, incluindo armazenamentos de chave-valor e bancos de dados. Inclua canários sintéticos e monitoramento de usuários reais em sua configuração de rastreamento completo a fim de medir as interações remotas com clientes e a

latência, para que você possa avaliar com precisão a performance de seus sistemas em relação aos seus objetivos e acordos de serviço.

Você pode usar o AWS X-Ray e os serviços de instrumentação de monitoramento de aplicações do Amazon CloudWatch para oferecer uma visão completa das solicitações à medida que elas passam por sua aplicação. O X-Ray coleta a telemetria da aplicação e permite que você a visualize e filtre em payloads, funções, rastreamentos, serviços, APIs e pode ser ativada para componentes do sistema com pouco ou nenhum código. O monitoramento de aplicações do CloudWatch inclui o ServiceLens para integrar seus rastreamentos a métricas, logs e alarmes. O monitoramento de aplicações do CloudWatch também inclui sintéticos a fim de monitorar seus endpoints e APIs, bem como monitoramento de usuários reais para instrumentar seus clientes de aplicações web.

Etapas da implementação

- Use o AWS X-Ray em todos os serviços nativos compatíveis, como Amazon S3, AWS Lambda e Amazon API Gateway. Esses serviços da AWS permitem ao X-Ray alternar a configuração usando a infraestrutura como código, AWS SDKs ou o AWS Management Console.
- Aplicações de instrumentos AWS Distro for Open Telemetry e X-Ray ou agentes de coleta de terceiros.
- Revise o Guia do desenvolvedor do AWS X-Ray para implementação específica da linguagem de programação. Essas seções da documentação detalham como instrumentar solicitações HTTP, consultas SQL e outros processos específicos de sua linguagem de programação de aplicações.
- Use o rastreamento do X-Ray para Canários sintéticos do Amazon CloudWatch e os Amazon CloudWatch RUM para analisar o caminho da solicitação de seu cliente de usuário final por meio de sua infraestrutura downstream da AWS.
- Configure métricas e alarmes do CloudWatch com base na integridade dos recursos e na telemetria canário para que as equipes sejam alertadas sobre problemas rapidamente e, depois, possam se aprofundar em rastreamentos e mapas de serviços com o ServiceLens.
- Habilite a integração do X-Ray a ferramentas de rastreamento de terceiros, como Datadog, New Relicou Dynatrace se você estiver usando ferramentas de terceiros para sua solução de rastreamento principal.

Recursos

Práticas recomendadas relacionadas:

- REL06-BP01 Monitorar todos os componentes da workload (geração)

- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

Documentos relacionados:

- O que é o AWS X-Ray?
- Amazon CloudWatch: Monitoramento de aplicações
- Depuração com o Amazon CloudWatch Synthetics e o AWS X-Ray
- A Amazon Builders' Library: instrumentação de sistemas distribuídos para visibilidade operacional
- Integrar o AWS X-Ray a outros serviços da AWS
- AWS Distro for OpenTelemetry e AWS X-Ray
- Amazon CloudWatch: uso do monitoramento sintético
- Amazon CloudWatch: usar o CloudWatch RUM
- Configurar o canário sintético do Amazon CloudWatch e o alarme do Amazon CloudWatch
- Disponibilidade e além: compreensão e melhoria da resiliência de sistemas distribuídos na AWS

Exemplos relacionados:

- Um workshop de observabilidade

Vídeos relacionados:

- AWS re:Invent 2022 - How to monitor applications across multiple accounts (re:Invent 2022: Como monitorar aplicações em várias contas)
- How to Monitor your AWS Applications (Como monitorar suas aplicações da AWS)

Ferramentas relacionadas:

- AWS X-Ray
- Amazon CloudWatch
- Amazon Route 53

CONFIABILIDADE 7. Como projetar sua workload 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. Notifique o pessoal apropriado para escalar manualmente o recurso ou inicie a automação para escalá-lo automaticamente.

A escala pode ser ajustada manualmente para a workload (por exemplo, alterando o número de instâncias do EC2 em um grupo do Auto Scaling ou modificando o throughput de uma tabela do DynamoDB por meio do AWS Management Console ou da AWS CLI). No entanto, a automação deve ser usada sempre que possível (consulte Usar automação ao obter ou escalar recursos).

Resultado desejado: as atividades de ajuste de escala (de forma automática ou manual) são iniciadas para restaurar a disponibilidade após a detecção de uma falha ou de uma degradação da experiência do cliente.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

Implemente a observabilidade e o monitoramento em todos os componentes da workload para monitorar a experiência do cliente e detectar falhas. Defina os procedimentos, manuais ou automatizados, que escalam os recursos necessários. Para obter mais informações, consulte REL11-BP01 Monitorar todos os componentes da workload para detectar falhas.

Etapas da implementação

- Definir os procedimentos, manuais ou automatizados, que escalam os recursos necessários.
  - Os procedimentos de ajuste de escala dependem de como os diferentes componentes da workload são projetados.
  - Eles também variam dependendo da tecnologia subjacente utilizada.
    - Os componentes que usam o AWS Auto Scaling podem utilizar planos de ajuste de escala para configurar um conjunto de instruções para escalar os recursos. Se você trabalha com o AWS CloudFormation ou adiciona tags aos recursos da AWS, poderá configurar planos de ajuste de escala para diferentes conjuntos de recursos por aplicação. O Auto Scaling fornece recomendações para estratégias de ajuste de escala personalizadas para cada recurso. Depois que o plano de ajuste de escala for criado, o Auto Scaling combinará os métodos

de ajuste de escala dinâmica e preditiva para compatibilidade com a estratégia de ajuste de escala. Para obter mais detalhes, consulte Como funcionam os planos de escalabilidade.

- O Amazon EC2 Auto Scaling verifica se você tem 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 do Auto Scaling. É possível especificar o número mínimo e máximo de instâncias em cada grupo do Auto Scaling, e o Amazon EC2 Auto Scaling garantirá que o grupo nunca fique abaixo ou acima desses limites. Para obter mais detalhes, consulte O que é o Amazon EC2 Auto Scaling?
- O ajuste de escala automático do Amazon DynamoDB usa o serviço Application Auto Scaling para ajustar dinamicamente a capacidade de throughput provisionado por você, em resposta a padrões de tráfego reais. Isso permite que uma tabela ou um índice secundário global aumente a capacidade provisionada de leitura e gravação para sustentar aumentos repentinos no tráfego, sem controle de utilização. Para obter mais detalhes, consulte Gerenciar a capacidade de throughput automaticamente com o Auto Scaling do DynamoDB.

Recursos

Práticas recomendadas relacionadas:

- REL07-BP01 Usar a automação ao obter ou escalar recursos
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

Documentos relacionados:

- AWS Auto Scaling: Como funcionam os planos de escalabilidade
- Gerenciar a capacidade de throughput automaticamente com o Auto Scaling do DynamoDB
- 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

CONFIABILIDADE 8. Como implementar uma alteração?

As alterações controladas são necessárias para implantar novas funcionalidades e garantir que as workloads 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 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 infraestrutura imutável

A infraestrutura imutável é um modelo que não requer atualizações, patches de segurança ou que alterações na configuração ocorram no local nas workloads de produção. Quando uma alteração é necessária, a arquitetura é criada em uma nova infraestrutura e implantada na produção.

Siga uma estratégia de implantação de infraestrutura imutável para aumentar a confiabilidade, a consistência e a reprodutibilidade nas implantações de workload.

Resultado desejado: com a infraestrutura imutável, nenhuma modificação no local é permitida para executar recursos de infraestrutura em uma workload. Em vez disso, quando uma alteração é necessária, um novo conjunto de recursos atualizados da infraestrutura que contém todas as alterações necessárias é implantado paralelamente aos recursos existentes. Essa implantação é validada automaticamente e, se bem-sucedida, o tráfego é gradualmente transferido para o novo conjunto de recursos.

Essa estratégia de implantação aplica-se a atualizações de software, patches de segurança, alterações na infraestrutura, atualizações de configuração e atualizações de aplicações, entre outros.

Antipadrões comuns:

- Implementação de mudanças no local em recursos da infraestrutura em execução.

Benefícios do estabelecimento desta prática recomendada:

- Maior consistência entre os ambientes: como não há diferenças nos recursos de infraestrutura entre os ambientes, a consistência aumenta e os testes são simplificados.
- Redução dos desvios da configuração: ao substituir os recursos da infraestrutura por uma configuração conhecida e controlada por versão, a infraestrutura é definida para um estado conhecido, testado e confiável, o que evita desvios de configuração.
- Implantações atômicas confiáveis: as implantações são concluídas com sucesso ou, do contrário, nada muda, aumentando a consistência e a confiabilidade no processo de implantação.
- Implantações simplificadas: as implantações são simplificadas porque não precisam oferecer suporte a atualizações. As atualizações são apenas novas implantações.
- Implantações mais seguras com processos de reversão e recuperação rápidos: as implantações são mais seguras porque a versão de trabalho anterior não é alterada. Você pode reverter para ele se forem detectados erros.
- Procedimento de segurança aprimorado: quando não se permitem alterações na infraestrutura, os mecanismos de acesso remoto (como o SSH) podem ser desativados. Isso reduz o vetor de ataque, melhorando o procedimento de segurança da organização.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

Automação

Ao definir uma estratégia de implantação de infraestrutura imutável, é recomendável usar a automação o máximo possível para aumentar a reprodutibilidade e minimizar a possibilidade de erro humano. Para obter mais detalhes, consulte REL08-BP05 Implantar alterações com automação e Automating safe, hands-off deployments.

Com a infraestrutura como código (IaC), as etapas de provisionamento, orquestração e implantação da infraestrutura são definidas de forma programática, descritiva e declarativa e armazenadas em um sistema de controle de origem. O uso da infraestrutura como código simplifica a automatização da implantação da infraestrutura e ajuda a obter a imutabilidade da infraestrutura.

Padrões de implantação

Quando uma mudança na workload é necessária, a estratégia de implantação da infraestrutura imutável exige que um novo conjunto de recursos da infraestrutura seja implantado, incluindo todas as alterações necessárias. É importante que esse novo conjunto de recursos siga um padrão de implantação que minimize o impacto sobre o usuário. Há duas estratégias principais para essa implantação:

Implantação canário: prática de direcionar um pequeno número de 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, será possível continuar a implantação a uma velocidade desejada e monitorar 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.

Implantação azul/verde: é semelhante à implantação canário, exceto que uma frota completa da aplicação é implantada 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 vez. No entanto, também é possível 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.

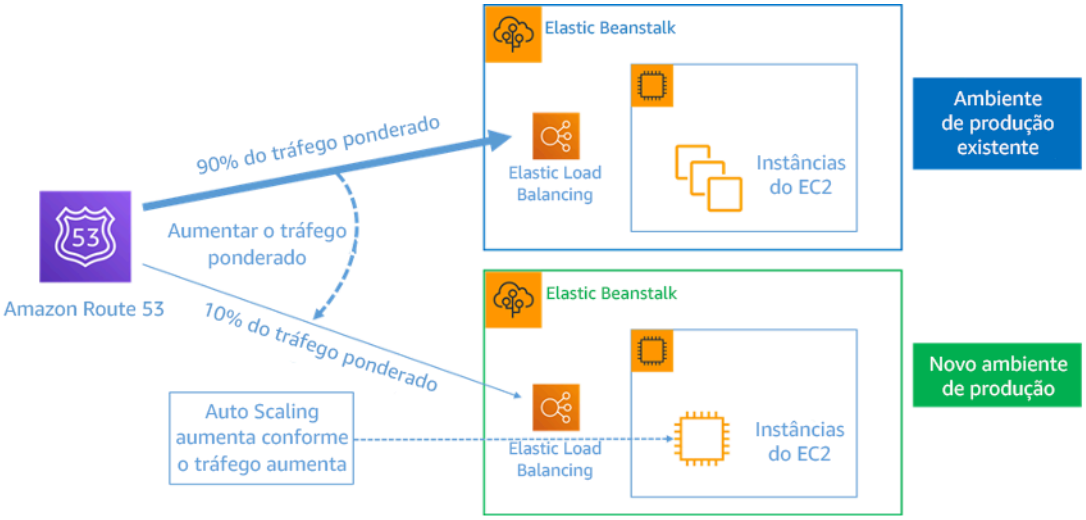


Figura 8: Implantação azul/verde com o AWS Elastic Beanstalk e o Amazon Route 53

Detecção de desvios

Define-se desvio como qualquer alteração que faça com que um recurso da infraestrutura tenha um estado ou uma configuração diferente do esperado. Alterações não gerenciadas da configuração, sejam de que tipo for, são contrárias ao conceito de infraestrutura imutável e devem ser detectadas e corrigidas para que a implementação da infraestrutura imutável seja bem-sucedida.

Etapas da implementação

- Proibir a modificação no local dos recursos de infraestrutura em execução.
  - É possível usar o AWS Identity and Access Management (IAM) para especificar quem ou o que pode acessar serviços e recursos na AWS, gerenciar as permissões refinadas centralmente e analisar o acesso para refinar as permissões em toda a AWS.
- Automatize a implantação dos recursos da infraestrutura para aumentar a reprodutibilidade e minimizar a possibilidade de erro humano.
  - Conforme descrito no whitepaper Introdução ao DevOps na AWS, a automação é uma referência dos serviços da AWS e é compatível com todos os serviços, recursos e ofertas.
  - Pré-fabricar a imagem de máquina da Amazon (AMI) pode acelerar o tempo para iniciá-la. O EC2 Image Builderé um serviço da AWS totalmente gerenciado que ajuda você a automatizar a criação, a manutenção, a validação, o compartilhamento e a implantação de AMIs personalizadas, seguras e atualizadas para Linux ou Windows.
  - Alguns dos serviços compatíveis com automação são:

    - O AWS Elastic Beanstalké um serviço para implantar e escalar rapidamente aplicações web desenvolvidas com Java, .NET, PHP, Node.js, Python, Ruby, Go e Docker em servidores conhecidos, como Apache, NGINX, Passenger e IIS.
    - O AWS Proton ajuda as equipes da plataforma a se conectarem e coordenarem todas as diferentes ferramentas de que as equipes de desenvolvimento precisam para provisionamento de infraestrutura, implantações de código, monitoramento e atualizações. O AWS Proton permite o provisionamento de infraestrutura como código automatizada e a implantação de aplicações sem servidor e baseadas em contêineres.
- A utilização da infraestrutura como código facilita a automatização da implantação da infraestrutura e ajuda a obter a imutabilidade da infraestrutura. A AWS fornece serviços que permitem a criação, a implantação e a manutenção da infraestrutura de forma programática, descritiva e declarativa.
    - O AWS CloudFormation ajuda os desenvolvedores a criar recursos da AWS de forma ordenada e previsível. Os recursos são escritos em arquivos de texto usando o formato JSON ou YAML. Os modelos exigem sintaxe e estrutura específicas que dependem dos tipos de recurso que estão sendo criados e gerenciados. Você cria os recursos em JSON ou YAML com qualquer editor de código, como o AWS Cloud9, e os insere em um sistema de controle de versão, e o CloudFormation cria os serviços especificados de maneira segura e repetível.
    - O AWS Serverless Application Model (AWS SAM) é uma estrutura de código aberto que você pode usar para criar aplicações sem servidor na AWS. O AWS SAM integra-se a outros serviços da AWS e é uma extensão do AWS CloudFormation.
    - O AWS Cloud Development Kit (AWS CDK) é um framework de desenvolvimento de software de código aberto para modelar e provisionar recursos da aplicação em nuvem usando linguagens de programação conhecidas. É possível usar o AWS CDK para modelar a infraestrutura de aplicações usando TypeScript, Python, Java e .NET. O AWS CDK usa o AWS CloudFormation em segundo plano para provisionar recursos de forma segura e repetível.
    - O AWS Cloud Control API apresenta um conjunto comum de APIs de criação, leitura, atualização, exclusão e lista (CRUDL) para ajudar os desenvolvedores a gerenciar a infraestrutura em nuvem de forma fácil e consistente. As APIs comuns do Cloud Control API permitem que os desenvolvedores gerenciem de maneira uniforme o ciclo de vida de serviços da AWS e de terceiros.
- Implemente padrões de implantação que minimizem o impacto no usuário.
    - Implantações canário:

    - Configurar uma implantação de versão canary do API Gateway
    - Create a pipeline with canary deployments for Amazon ECS using AWS App Mesh
  - Implantações azul/verde: o whitepaper Blue/Green Deployments on AWS descreve exemplos de técnicas para implementar estratégias de implantação azul/verde.
- Detecte variações da configuração ou do estado. Para obter mais detalhes, consulteDetectar alterações de configuração não gerenciadas em pilhas e recursos.

Recursos

Práticas recomendadas relacionadas:

- REL08-BP05 Implantar alterações com automação

Documentos relacionados:

- Automatizar uma implantação prática e sem intervenção manual
- Leveraging AWS CloudFormation to create an immutable infrastructure at Nubank
- Infraestrutura como código
- Implementing an alarm to automatically detect drift in AWS CloudFormation stacks

Vídeos relacionados:

- AWS re:Invent 2020: Reliability, consistency, and confidence through immutability

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

- CONFIABILIDADE 9. Como fazer backup dos dados?
- CONFIABILIDADE 10. Como usar o isolamento de falhas para proteger sua workload?
- CONFIABILIDADE 11. Como projetar a workload para resistir a falhas de componentes?
- CONFIABILIDADE 12. Como testar a confiabilidade?
- CONFIABILIDADE 13. Como planejar a recuperação de desastres (DR)?

CONFIABILIDADE 9. Como fazer 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 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

Compreenda e use os recursos de backup dos serviços e recursos de dados usados pela workload. A maioria dos serviços oferece recursos para fazer backup dos dados da workload.

Resultado desejado: as fontes de dados foram identificadas e classificadas com base na criticidade. Depois, estabeleça uma estratégia de recuperação de dados com base no RPO. A estratégia envolve fazer backup dessas fontes de dados ou poder 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: básica

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 e 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 é estabelecida: alto

Orientações para a implementação

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), permitindo restaurar um backup a qualquer momento até cinco minutos ou menos, antes da hora atual. Muitos serviços da AWS permitem copiar backups para outra Região da AWS. O AWS Backup é uma ferramenta que permite centralizar e automatizar a proteção de dados nos serviços da AWS. O AWS Elastic Disaster Recovery permite copiar workloads de servidor completas e manter uma proteção de dados contínua de um ambiente on-premises, entre zonas de disponibilidade ou entre regiões, com um objetivo de ponto de recuperação (RPO) medido em segundos.

O Amazon S3 pode ser usado como um 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.

E possível fazer backup de dados on-premises na Nuvem AWS usando o AWS Storage Gateway ou o AWS DataSync. Os buckets do Amazon S3 podem ser usados para armazenar esses dados na AWS. O Amazon S3 oferece vários níveis de armazenamento, como Amazon S3 Glacier ou S3 Glacier Deep Archive para reduzir os custos do armazenamento de dados.

Você pode atender às necessidades de recuperação de dados reproduzindo os dados de outras fontes. Por exemplo, os nós de réplica do Amazon ElastiCache ou as réplicas de leitura do Amazon RDS poderiam ser usadas para reproduzir dados caso os primários sejam perdidos. Em casos em que fontes como essa podem ser usadas para atender ao objetivo de ponto de recuperação (RPO) e objetivo de tempo de recuperação (RTO), pode ser que você não precise de um backup. Outro exemplo, se estiver trabalhando com o Amazon EMR, pode não ser necessário fazer backup do armazenamento de dados HDFS, contanto que você possa reproduzir os dados no Amazon EMR pelo Amazon 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.

Etapas da implementação

1. Identifique todas as fontes de dados para a workload. Os dados podem ser armazenados em vários recursos, como bancos de dados, volumes, sistemas de arquivos, sistemas de registro e armazenamento de objetos. Consulte a seção Recursos para encontrar Documentos relacionados sobre diferentes serviços da AWS onde os dados são armazenados e o recurso de backup que esses serviços 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 serviços de terceiros para criar backups dos dados. O AWS Backup é um serviço gerenciado que permite criar backups de várias fontes de dados na AWS. O AWS Elastic Disaster Recovery lida com a replicação de dados automáticos de subsegundos em uma Região da AWS. A maioria dos serviços da AWS também possui recursos nativos para criar backups. O AWS Marketplace tem muitas soluções que também fornecem esses recursos. Consulte os 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. Às 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 uma frequência 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 do 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 cumprir os objetivos de recuperação

Documentos relacionados:

- O que é o Reachability Analyzer?
- What is AWS DataSync? (O que é o AWS Data Sync?)
- What is Volume Gateway? (O que é o Gateway de Volumes?)
- Parceiro do APN: parceiros que podem ajudar com o backup
- AWS Marketplace: products that can be used for backup (AWS Marketplace: produtos que podem ser usados para backup)
- Snapshots do Amazon EBS
- Backing Up Amazon EFS (Fazer backup do Amazon EFS)
- Backing up Amazon FSx for Windows File Server (Fazer backup do Amazon FSx para Windows File Server)
- Backup e restauração para o ElastiCache for Redis
- Creating a DB Cluster Snapshot in Neptune (Criar um snapshot do cluster de banco de dados no Neptune)
- Criar um snapshot do banco de dados
- Creating an EventBridge Rule That Triggers on a Schedule (Criar uma regra do EventBridge que é acionada de acordo com uma programação)
- Replicação entre regiões com o Amazon S3
- EFS para EFS AWS Backup
- Exporting Log Data to Amazon S3 (Exportação de dados de log para o Amazon S3)
- Gerenciamento do ciclo de vida de objetos
- Usar backup e restauração sob demanda para o DynamoDB
- Recuperação pontual para DynamoDB
- Criação de snapshots de índices no Amazon OpenSearch Service

- O que é o Reachability Analyzer?

Vídeos relacionados:

- AWS re:Invent 2021: Backup, disaster recovery, and ransomware protection with AWS (Backup, recuperação de desastres e proteção contra ransomware com a AWS)
- AWS Backup Demo: Cross-Account and Cross-Region Backup (Demonstração: Backup entre contas e entre regiões)
- AWS re:Invent 2019: Deep dive on AWS Backup, ft. Rackspace (STG341)

Exemplos relacionados:

- Well-Architected Lab - Implementing Bi-Directional Cross-Region Replication (CRR) for Amazon S3 (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
- Well-Architected Lab - Backup and Restore with Failback for Analytics Workload (Laboratório do Well-Architected: backup e restauração com failback para workload do Analytics)
- Well-Architected Lab - Disaster Recovery - Backup and Restore (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. Use a criptografia para prevenir e detectar se a integridade dos dados de backups está comprometida.

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 é estabelecida: alto

Orientações para a implementação

Controle e detecte o acesso a backups usando autenticação e autorização, como o AWS Identity and Access Management (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 do lado do servidor, o Amazon S3 aceita os objetos como dados não criptografados e, depois, 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. Ao usar o AWS Elastic Disaster Recovery, todos os dados em trânsito e em repouso são criptografados. Com o Elastic Disaster Recovery, os dados em repouso podem ser criptografados usando a chave de criptografia de volume padrão do Amazon EBS ou uma chave personalizada gerenciada pelo cliente.

Etapas da implementação

1. Use a criptografia em cada um dos seus armazenamentos de dados. Se os dados de origem forem criptografados, o backup também será.
    - Use criptografia no Amazon RDS. Você pode configurar a criptografia em repouso usando o AWS Key Management Service ao criar uma instância do RDS.
    - Use criptografia em volumes do Amazon EBS. Você pode configurar a criptografia padrão ou especificar uma chave exclusiva após a criação do volume.
    - Use a criptografia do Amazon DynamoDB necessária. O DynamoDB criptografa todos os dados em repouso. Você pode usar uma chave do AWS KMS de propriedade da AWS ou uma chave do KMS gerenciada pela AWS, especificando uma chave armazenada na sua conta.
    - Criptografe seus dados armazenados no Amazon EFS. Configure a criptografia ao criar seu sistema de arquivos.
    - 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.

- Escolha se deseja usar a critptografia do Amazon EBS para o Elastic Disaster Recovery padrão ou personalizada. Essa opção criptografa os dados em repouso replicados nos discos da sub-rede da área de preparação e os discos replicados.

2. Implemente permissões de privilégio mínimo para acessar seus backups. Siga as práticas recomendadas para limitar o acesso aos backups, aos snapshots e às réplicas de acordo com as práticas recomendadas de segurança.

Recursos

Documentos relacionados:

- AWS Marketplace: products that can be used for backup (AWS Marketplace: produtos que podem ser usados para backup)
- Criptografia do Amazon EBS
- Amazon S3: proteção de dados usando criptografia
- Replicar objetos criados com criptografia do lado do servidor (SSE) usando chaves do AWS KMS
- Criptografia em repouso do DynamoDB
- Criptografar recursos do Amazon RDS
- Encrypting Data and Metadata in Amazon EFS (Criptografia de dados e metadados no Amazon EFS)
- Encryption for Backups in AWS (Criptografia para backups na AWS)
- Gerenciamento de tabelas criptografadas
- Pilar Segurança: AWS Well-Architected Framework
- O que é o AWS Elastic Disaster Recovery?

Exemplos relacionados:

- Well-Architected Lab - Implementing Bi-Directional Cross-Region Replication (CRR) for Amazon S3 (Laboratório do Well-Architected: implementação da replicação bidirecional entre regiões (CRR) para o Amazon S3)

REL09-BP03 Realizar backup de dados automaticamente

Configure os backups para serem feitos automaticamente com base em uma programação periódica informada pelo objetivo de ponto de recuperação (RPO) ou de acordo com alterações no conjunto de

dados. É necessário fazer frequentemente o backup automático de conjuntos de dados críticos com requisitos de baixa perda de dados, enquanto o backup de dados menos críticos, em que alguma perda é aceitável, pode ser feito com menos frequência.

Resultado desejado: um processo automatizado que cria backups de fontes de dados em uma frequência estabelecida.

Antipadrões comuns:

- Fazer backups manualmente.
- Usar recursos que têm o recurso de backup, mas não incluir o backup em sua automação.

Benefícios do estabelecimento dessa 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 risco exposto se esta prática recomendada não é estabelecida: médio

Orientações para a implementação

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 RDS e 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 permite programar backups periódicos, até de hora em hora.

O Amazon RDS e o Amazon DynamoDB oferecem backup contínuo com recuperação a um ponto anterior no tempo. O versionamento do Amazon S3, quando habilitado, é automático. O Amazon Data Lifecycle Manager pode ser usado para automatizar a criação, a cópia e a exclusão de snapshots do Amazon EBS. Ele também pode automatizar a criação, a cópia, a suspensão e o cancelamento do registro de imagens de máquina da Amazon (AMIs) com base no Amazon EBS e seus snapshots subjacentes do Amazon EBS.

O AWS Elastic Disaster Recovery fornece replicação contínua no nível de bloco do ambiente de origem (on-premises ou a AWS) para a região de recuperação de destino. Os snapshots do Amazon EBS de um ponto anterior no tempo são criados automaticamente e gerenciados pelo serviço.

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.

Etapas da implementação

1. Identifique fontes de dados que estão sendo copiados manualmente. Para obter mais detalhes, consulte REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes.
2. Determine o RPO da workload. Para obter mais detalhes, consulte REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados.
3. Use uma solução de backup automático ou um serviço gerenciado. O AWS Backup é um serviço totalmente gerenciado que facilita centralizar e automatizar a proteção de dados nos serviços da AWS, na nuvem e no ambiente on-premises. O uso de planos no AWS Backup permite a criação de regras que definem os recursos para backup e a frequência com que esses backups devem ser criados. A frequência deve ser informada pelo RPO estabelecido na Etapa 2. Para obter orientações práticas sobre como criar backups automáticos usando o AWS Backup, consulte Testing Backup and Restore of Data (Teste de backup e restauração de dados). 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 incompatíveis com 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.

Nível de esforço do plano de implementação: baixo

Recursos

Documentos relacionados:

- Parceiro do APN: parceiros que podem ajudar com o backup
- AWS Marketplace: products that can be used for backup (AWS Marketplace: produtos que podem ser usados para backup)
- Creating an EventBridge Rule That Triggers on a Schedule (Criar uma regra do EventBridge que é acionada de acordo com uma programação)
- O que é o AWS Backup?
- O que é o Reachability Analyzer?
- O que é o AWS Elastic Disaster Recovery?

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 de tempo de recuperação (RTO) e de ponto de recuperação (RPO).

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 dessa 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 testes também garantem que a restauração e a recuperação sejam possíveis de acordo com o RTO da workload e qualquer perda de dados se enquadre no RPO da workload.

Nível de risco exposto se esta prática recomendada não é 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 verificar se todos os dados estão disponíveis, não corrompidos, acessíveis e garantir que toda a perda de dados se enquadre no RPO da workload. Eles também podem ajudar a verificar se os mecanismos de recuperação são rápidos o suficiente para acomodar o RTO da workload.

Ao usar a AWS, você pode criar um ambiente de teste e restaurar os 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.

Se todos os dados estiverem disponíveis, não corrompidos, acessíveis e qualquer perda de dados está de acordo com o RPO da workload. Eles também podem ajudar a verificar se os mecanismos de recuperação são rápidos o suficiente para acomodar o RTO da workload.

O AWS Elastic Disaster Recovery oferece snapshots contínuos de recuperação a um ponto anterior no tempo de volumes do Amazon EBS. Como os servidores de origem são replicados, os estados de

um ponto anterior no tempo são registrados ao longo do tempo com base na política configurada. O Elastic Disaster Recovery ajuda a verificar a integridade desses snapshots iniciando instâncias para fins de teste e simulação sem redirecionar o tráfego.

Etapas da implementação

1. Identifique fontes de dados das quais está sendo feito backup no momento e onde esses backups estão sendo armazenados. Para obter orientações de implementação, consulte REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes.
2. Estabeleça critérios para a validação de dados a 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 o RTO e o RPO para restaurar os dados com base na criticidade deles. Para obter orientações de implementação, consulte REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados.
4. Avalie sua capacidade de recuperação. Revise sua estratégia de backup e de restauração para entender se ela pode cumprir o RTO e o RPO e ajuste a estratégia conforme necessário. Usando o AWS Resilience Hub, você pode executar uma avaliação da workload. Essa avaliação analisa a configuração da aplicação em relação à política de resiliência e relata se as metas de RTO e RPO podem ser cumpridas.
5. Faça uma restauração de teste com os 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 pode ser tão simples quanto restaurar o backup em um novo recurso. Se você usar o AWS Elastic Disaster Recovery, poderá iniciar uma simulação de recuperação.
6. Valide a recuperação de dados do recurso restaurado com base nos critérios estabelecidos anteriormente para validação dos dados. Os dados restaurados e recuperados contêm o registro ou item mais recente no momento do backup? Esses dados se enquadram no RPO da workload?
7. Calcule o tempo necessário para a restauração e recuperação e compare-o com o RTO estabelecido. Esse processo se enquadra no RTO da workload? Por exemplo, compare o carimbo

de data/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 carimbo de data/hora e essas informações estão disponíveis no AWS CloudTrail. Embora essas informações possam fornecer detalhes sobre o início do processo de restauração, o carimbo final de data/hora da conclusão da validação deve ser registrado pela lógica de validação. Se você usar um processo automático, serviços como o Amazon DynamoDB poderão ser usados para armazenar essas informações. Além disso, muitos serviços da AWS oferecem um histórico de eventos que fornece informações sobre a data e hora em que determinadas ações ocorreram. No AWS Backup, as ações de backup e restauração são chamadas de trabalhos, e esses trabalhos contêm informações de data e hora como parte dos metadados, que podem ser usados para calcular o tempo necessário para restauração e recuperação.

8. Informe 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, serviços como o Amazon Simple Notification Service (Amazon SNS) podem ser usados para enviar notificações por push como e-mail ou SMS às partes interessadas. Essas mensagens também podem ser publicadas em aplicações 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 esse 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 a validação da recuperação de dados com o AWS Backup. Além disso, este laboratório do Well-Architected fornece uma experiência prática de como automatizar várias das etapas aqui.

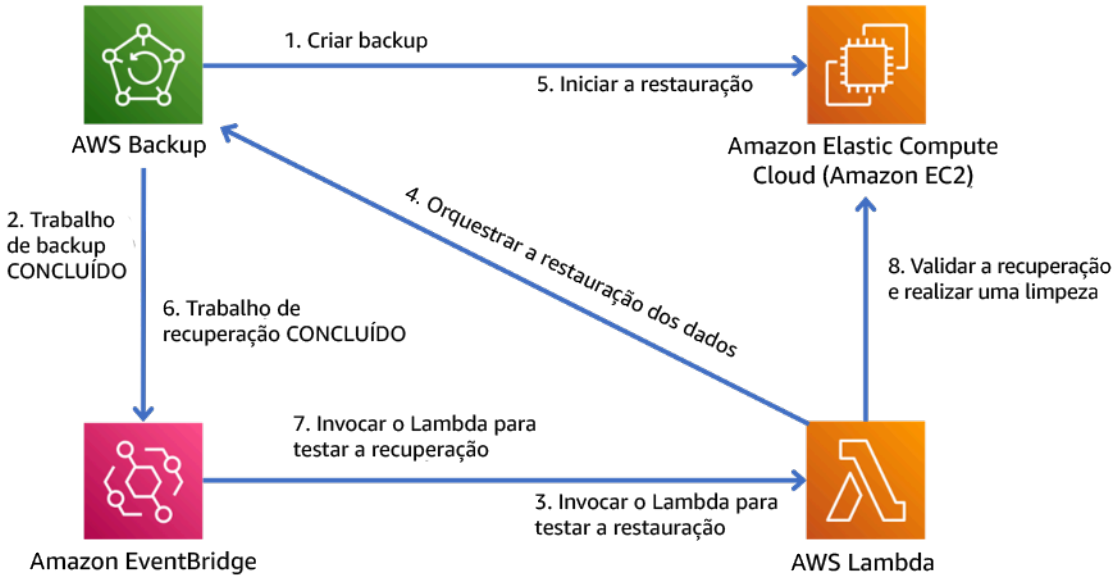


Figura 9. Um processo de backup e restauração automatizado

Nível de esforço do plano de implementação: moderado a alto, dependendo da complexidade dos critérios de validação.

Recursos

Documentos relacionados:

- Automate data recovery validation with AWS Backup (Automatizar validação de recuperação de dados com o AWS Backup)
- Parceiro do APN: parceiros que podem ajudar com o backup
- AWS Marketplace: products that can be used for backup (AWS Marketplace: produtos que podem ser usados para backup)
- Creating an EventBridge Rule That Triggers on a Schedule (Criar uma regra do EventBridge que é acionada de acordo com uma programação)
- Usar backup e restauração sob demanda para o DynamoDB
- O que é o AWS Backup?
- O que é o Reachability Analyzer?
- O que é o Reachability Analyzer?
- AWS Elastic Disaster Recovery

Exemplos relacionados:

- Laboratório do Well-Architected: teste de backup e restauração de dados

CONFIABILIDADE 10. Como usar o isolamento de falhas para proteger sua workload?

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.

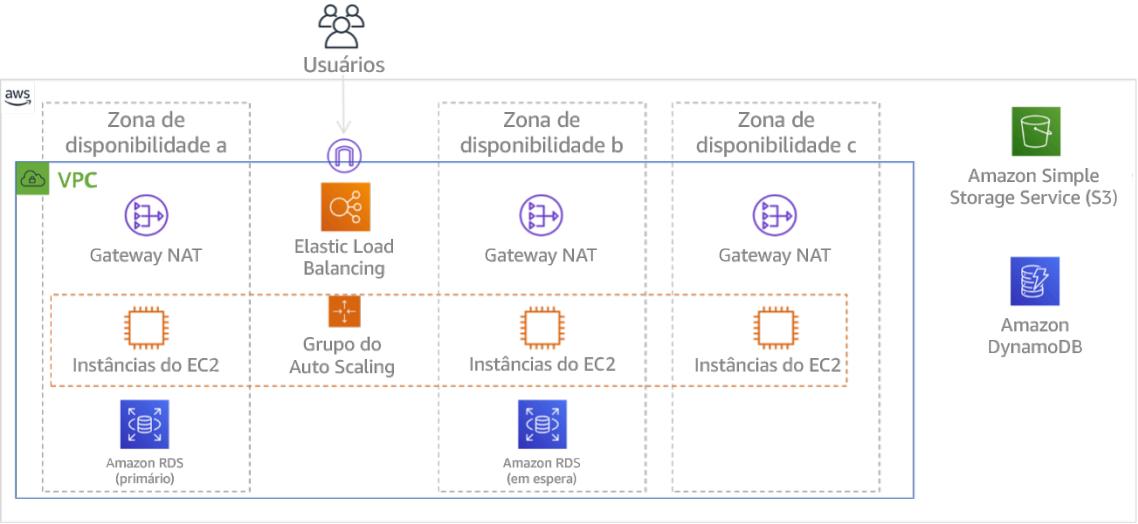


Figura 9: arquitetura multicamadas implantada em três Zonas de disponibilidade. Observe que o Amazon S3 e o Amazon DynamoDB são sempre multi-AZ automaticamente. O ELB também é implantado em todas as três zonas.

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.

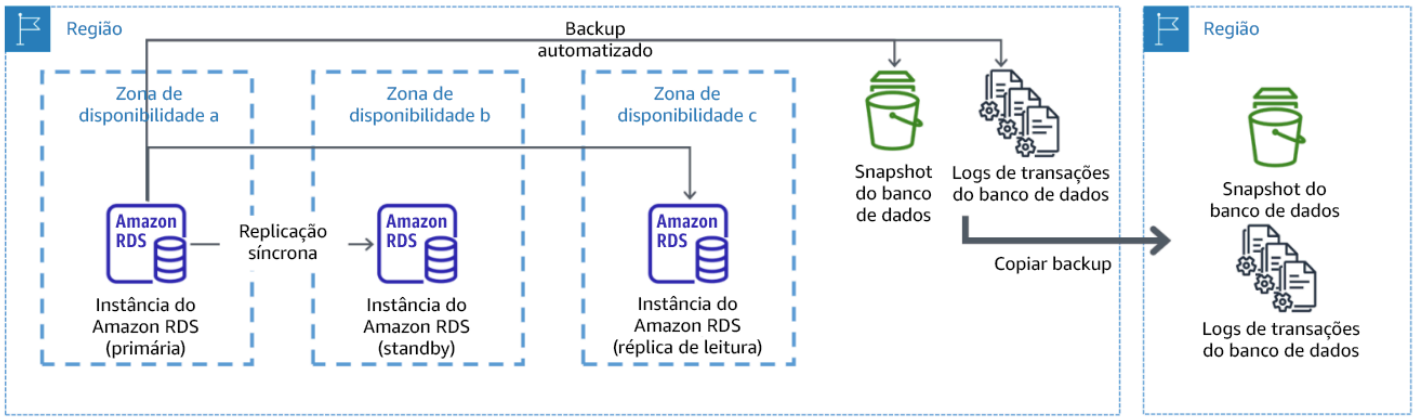


Figura 10: Uma implantação de banco de dados multi-AZ resiliente com backup para outra região da AWS

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 cumprir os 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 datacenter on-premises, você deverá implementar capacidade suficiente para fazer uma recompilação completa da workload em conformidade com os objetivos de recuperação definidos.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientações para a implementação

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. Usando o EMR File System (EMRFS), os dados no EMR podem ser armazenados no Amazon S3, que, por sua vez, 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.

Para workloads com estado baseadas em servidor e implantadas em um datacenter on-premises, é possível usar o AWS Elastic Disaster Recovery para proteger as workloads na AWS. Se você já estiver hospedado na AWS, poderá usar o Elastic Disaster Recovery para proteger a workload em uma zona de disponibilidade ou região alternativa. O Elastic Disaster Recovery usa a replicação contínua no nível de bloco para uma área de preparação leve visando fornecer recuperação rápida e confiável de aplicações on-premises e na nuvem.

Etapas da implementação

1. 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 do Amazon EC2 Auto Scaling para workloads de contêiner e instâncias que não têm requisitos para um endereço IP de instância única, endereço IP privado, endereço IP elástico e metadados da instância.
     - Os dados do usuário do modelo de execução podem ser usados para implementar uma automação que pode recuperar automaticamente a maioria das workloads.
   - Use a recuperação automática de instâncias do Amazon EC2 para workloads que exijam um único endereço de ID da instância, endereço IP privado, endereço IP elástico e metadados da 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 do ciclo de vida da instância do Amazon EC2 ou os eventos do Amazon ECS para automatizar a autorreparação, em que não seja possível usar a escalabilidade automática ou a recuperação do EC2.
     - Use os eventos para chamar a automação que recuperará seu componente de acordo com a lógica do processo necessária.
   - Proteja workloads com estado que são limitadas a um único local usando o AWS Elastic Disaster Recovery.

Recursos

Documentos relacionados:

- Eventos do Amazon ECS

- Ganchos do ciclo de vida do Amazon EC2 Auto Scaling
- Recupere sua instância.
- Escalabilidade automática do serviço
- O que é o Amazon EC2 Auto Scaling?
- AWS Elastic Disaster Recovery

REL10-BP04 Usar arquiteturas de anteparo para limitar o escopo de impacto

Implemente arquiteturas de anteparo (também chamadas de arquiteturas baseadas em células) para restringir o efeito ou a falha em uma workload a um número limitado de componentes.

Resultado desejado: uma arquitetura baseada em células usa várias instâncias isoladas de uma workload, em que cada instância é considerada uma célula. Cada célula é independente, não compartilha o estado com outras células e processa um subconjunto das solicitações gerais da workload. Isso reduz o possível impacto de uma falha, como uma atualização de software incorreta, a uma célula individual e às solicitações que ela está processando. Se uma workload usa 10 células para atender a 100 solicitações, quando ocorre uma falha, 90% das solicitações gerais não seriam afetadas pela falha.

Antipadrões comuns:

- Permitir que as células cresçam sem limites.
- Aplicar implantações ou atualizações de código a todas as células ao mesmo tempo.
- Compartilhar o estado ou os componentes entre as células (com a exceção da camada do roteador).
- Adicionar negócios complexos ou rotear lógica para a camada do roteador.
- Não minimizar as interações entre as células.

Benefícios do estabelecimento dessa prática recomendada: com arquiteturas baseadas em células, muitos tipos comuns de falhas são contidas na própria célula, fornecendo isolamento de falhas adicional. Esses limites de falhas podem fornecer resiliência com relação a tipos de falha que, de outra maneira, seriam difíceis de conter, como implantações de código sem êxito ou solicitações corrompidas ou que acionam um modo de falha específico (também conhecidas como solicitações com conteúdo malicioso).

Orientação de implementação

Em um navio, as anteparas garantem que uma ruptura no casco seja contida em uma seção do casco. Em sistemas complexos, esse padrão costuma ser replicado para permitir o isolamento de falhas. Os limites isolados de falhas restringem o efeito de uma falha em uma workload 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. Na AWS, os clientes podem usar várias zonas de disponibilidade e regiões para fornecer o isolamento de falhas, mas o conceito do isolamento de falhas também pode ser estendido à arquitetura da workload.

A workload geral é composta por células particionadas por uma chave de partição. Essa chave precisa se alinhar à granularidade do serviço, ou da maneira natural que a workload de um serviço pode ser subdividida em interações mínimas entre células. Exemplos de chaves de partição são ID de cliente, ID de recurso ou qualquer outro parâmetro facilmente acessível na maioria das chamadas de API. Uma camada de roteamento de célula distribui solicitações a células individuais com base na chave de partição e apresenta um único endpoint aos clientes.

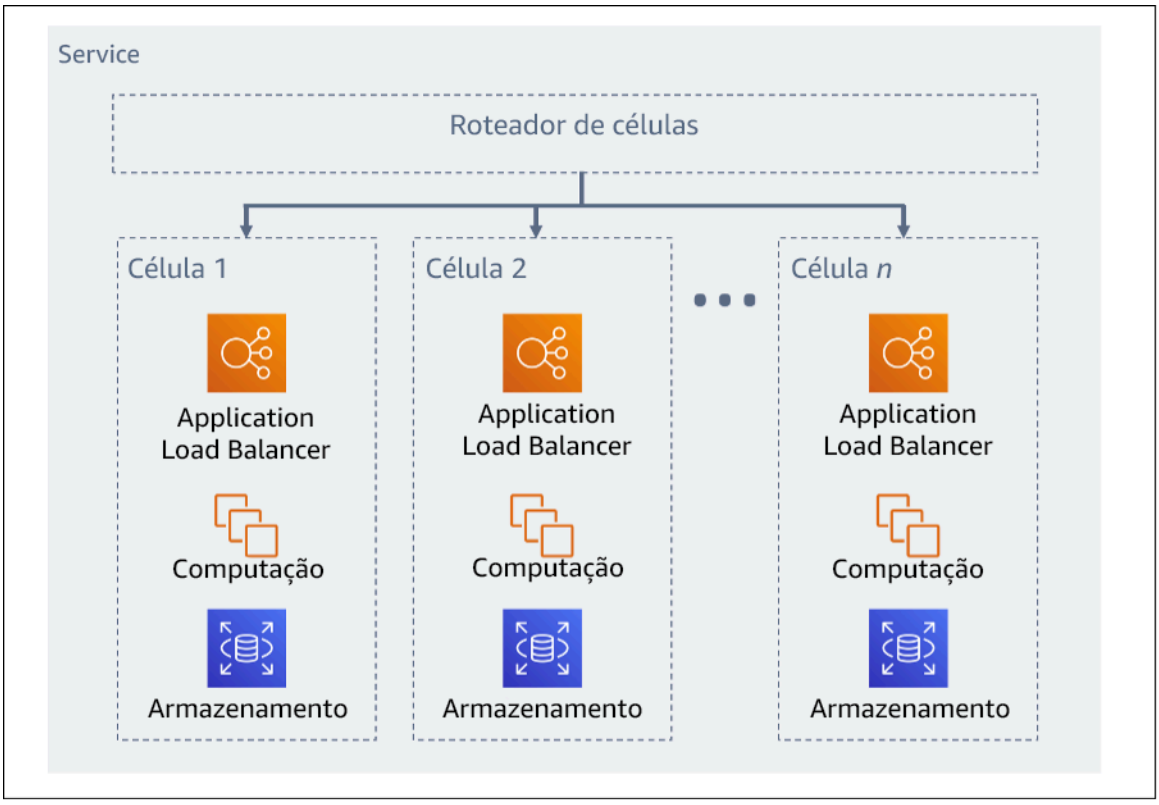


Figura 11: arquitetura baseada em células

Etapas da implementação

Ao projetar uma arquitetura baseada em células, há várias considerações de design a levar em conta:

1. Chave de partição: deve-se dedicar uma consideração especial ao escolher a chave de partição.
   - Ela precisa se alinhar à granularidade do serviço, ou da maneira natural que a workload de um serviço pode ser subdividida em interações mínimas entre células. Dentre os exemplos estão o `ID de cliente` ou `ID de recurso`.
   - A chave de partição deve estar disponível em todas as solicitações, seja diretamente ou de uma maneira que possa ser facilmente inferida de forma determinística por outros parâmetros.
2. Mapeamento de células persistentes: os serviços de upstream só devem interagir com uma única célula pelo ciclo de vida dos recursos.
   - Dependendo da workload, uma estratégia de migração de células pode ser necessária para migrar os dados de uma célula para outra. Um possível cenário de quando é necessário fazer uma migração de célula seria quando um usuário ou recurso específico na workload fica grande demais e exige uma célula dedicada.
   - As células não devem compartilhar estado ou componentes entre si.
   - Consequentemente, as interações entre as células devem ser evitadas e mantidas no mínimo, já que elas podem criar dependências entre as células e, assim, reduzir as melhorias do isolamento de falhas.
3. Camada do roteador: a camada do roteador é um componente compartilhado entre células e, portanto, não pode seguir a mesma estratégia de compartimentalização das células.
   - É recomendável que a camada do roteador distribua as solicitações para células individuais usando um algoritmo de mapeamento de partição de maneira computacionalmente eficiente, como combinando funções de hash criptográficas e aritmética modular para mapear chaves de partição a células.
   - Para evitar impactos em várias células, a camada de roteamento deve permanecer o mais simples e horizontalmente escalável possível, o que exige evitar uma lógica empresarial complexa nessa camada. Isso tem o benefício adicional de facilitar a compreensão de seu comportamento esperado em todos os momentos, permitindo uma capacidade de testes completa. Conforme explicado por Colm MacCárthaigh em Reliability, constant work, and a good cup of coffee (Confiabilidade, trabalho constante e uma boa xícara de café), designs simples e padrões de trabalho constantes produzem sistemas confiáveis e reduzem a antifragilidade.
4. Tamanho da célula: as células devem ter um tamanho máximo e não devem ter permissão para crescer além disso.

- O tamanho máximo deve ser identificado com a realização de testes completos, até que os pontos de ruptura sejam atingidos e as margens operacionais seguras sejam estabelecidas. Para obter mais detalhes sobre como implementar práticas de testes, consulte REL07-BP04 Fazer o teste de carga da sua workload
- A workload geral deve crescer com a adição de mais células, permitindo que a workload seja escalada com aumentos na demanda.

5. Estratégias de várias zonas de disponibilidade ou várias regiões: várias camadas de resiliência devem ser utilizadas para a proteção contra diferentes domínios de falha.
   - Para resiliência, você deve usar uma abordagem que crie 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 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. Para obter mais detalhes, consulte REL10-BP02 Escolher os locais apropriados para sua implantação de vários locais.
6. Implantação de código: uma estratégia de implantação de código escalonada deve ter preferência com relação à implantação de alterações de código em todas as células ao mesmo tempo.
   - Isso ajudará a reduzir a possibilidade de falhas em várias células devido a uma implantação incorreta ou a erro humano. Para obter mais detalhes, consulte Automating safe, hands-off deployment (Automatizar uma implantação prática e segura).

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

Recursos

Práticas recomendadas relacionadas:

- REL07-BP04 Fazer o teste de carga da sua workload
- REL10-BP02 Escolher os locais apropriados para sua implantação de vários locais

Documentos relacionados:

- Reliability, constant work, and a good cup of coffee (Confiabilidade, trabalho constante e uma boa xícara de café)
- AWS and Compartmentalization (AWS e compartimentalização)
- isolamento de carga de trabalho usando fragmentos aleatórios
- Automating safe, hands-off deployment (Automatizar uma implantação prática e segura)

Vídeos relacionados:

- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (AWS re:Invent 2018: fechar ciclos e abrir mentes: como controlar sistemas, sejam grandes ou pequenos)
- AWS re:Invent 2018: como a AWS reduz o impacto das falhas (ARC338)
- Fragmentação aleatória: AWS re:Invent 2019: apresentação da Amazon Builders' Library (DOP328)
- AWS Summit ANZ 2021: tudo falha o tempo todo: como criar o design visando a resiliência

Exemplos relacionados:

- Laboratório do Well-Architected: isolamento de falhas com fragmentação aleatória

CONFIABILIDADE 11. Como projetar a workload 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 estabilidade estática para evitar o comportamento bimodal
- REL11-BP06 Enviar notificações quando os eventos afetarem a disponibilidade

- REL11-BP07 Arquitetar o produto para cumprir as metas de disponibilidade e os acordos de nível de serviço (SLAs) de tempo de atividade

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 falhas ou degradações assim que elas ocorrerem. Monitore os indicadores-chave de performance (KPIs) 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 de entregar valor empresarial, portanto, os indicadores-chave de performance (KPIs) que medem isso precisam fazer parte da sua estratégia de detecção e remediação.

Resultado desejado: Os componentes essenciais de uma workload são monitorados de forma independente para detectar e alertar sobre falhas quando e onde elas acontecem.

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 objetivo de tempo de recuperação (RTO)
- Somente as interfaces da workload voltadas para o cliente são monitoradas ativamente.
- Coleta apenas das métricas técnicas, não das métricas de função de negócios.
- Não há métricas que medem a experiência do usuário da workload.
- São criados monitores em excesso.

Benefícios de estabelecer esta prática recomendada: O monitoramento adequado de todas as camadas permite reduzir o tempo de recuperação ao reduzir o tempo de detecção.

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

Orientação para implementação

Identifique todas as workloads que serão analisadas para monitoramento. Depois de identificar todos os componentes da workload que precisarão ser monitorados, você precisará determinar o intervalo

de monitoramento. O intervalo de monitoramento terá um impacto direto na rapidez com que a recuperação pode ser iniciada com base no tempo necessário para detectar uma falha. O tempo médio de detecção (MTTD) é a quantidade de tempo entre a ocorrência de uma falha e o início das operações de reparo. A lista de serviços deve ser extensa e completa.

O monitoramento deve abranger todas as camadas da pilha de aplicações, incluindo aplicação, plataforma, infraestrutura e rede.

Sua estratégia de monitoramento deve considerar o impacto de falhas cinzentas. Para obter mais detalhes sobre falhas cinzentas, consulte Falhas cinzentas no whitepaper Advanced Multi-AZ Resilience Patterns (Padrões avançados de resiliência Multi-AZ).

Etapas da implementação

- 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 objetivo de tempo de recuperação (RTO) para determinar a frequência da coleta.
- Configure o monitoramento detalhado de componentes e serviços gerenciados.
  - Determine se o monitoramento detalhado para instâncias do EC2 e Auto Scaling são necessários. O monitoramento detalhado fornece métricas de intervalo de um minuto, e o monitoramento padrão fornece métricas de intervalo de cinco minutos.
  - Determine se o monitoramento aprimorado para RDS é necessário. O monitoramento aprimorado usa um agente nas instâncias do RDS para obter informações úteis sobre processos ou threads diferentes.
  - Determine os requisitos de monitoramento de componentes essenciais sem servidor para Lambda, o API Gateway, o Amazon EKS, o Amazon ECS, e todos os tipos de balanceadores de carga.
  - Determine os requisitos de monitoramento dos componentes de armazenamento para Amazon S3, o Amazon FSx, o Amazon EFSe o Amazon EBS.
- Crie métricas personalizadas para medir os indicadores-chave de performance (KPIs) do negócio. As workloads implementam as principais funções empresariais, que devem ser usadas como KPIs para ajudar a identificar quando ocorre um problema indireto.
- Utilize os canários de usuário para monitorar a experiência do usuário e verificar se há falhas. O teste de transação sintética (também conhecido como “teste canário”, que não deve ser confundido com “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 workload de diversos locais remotos.

- Crie métricas personalizadas que rastreiem 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.
- Defina alarmes para detectar quando uma parte da workload não estiver funcionando corretamente e indicar quando deve ser feita a escalabilidade automática dos recursos. Os alarmes podem ser exibidos visualmente em painéis, enviar alertas pelo Amazon SNS ou por e-mail e trabalhar com o Auto Scaling para aumentar ou reduzir a escala dos recursos da workload.
- Crie painéis para visualizar suas 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.
- Crie monitoramento de rastreamento distribuído para seus serviços. Com o monitoramento distribuído, você compreende como está a performance de sua aplicação e seus serviços subjacentes para identificar e solucionar a causa principal de problemas e erros de performance.
- Crie painéis de sistemas de monitoramento (usando CloudWatch ou X-Ray) e coleta de dados em uma Região e conta separadas.
- Crie integração para o monitoramento do Amazon Health Aware para permitir a visibilidade de monitoramento de recursos da AWS que possam ter degradações. Para workloads essenciais aos negócios, essa solução fornece acesso a alertas proativos e em tempo real para serviços da AWS.

Recursos

Práticas recomendadas relacionadas:

- Definição de disponibilidade
- REL11-BP06 Enviar notificações quando os eventos afetarem a disponibilidade

Documentos relacionados:

- Amazon CloudWatch Synthetics permite criar canários de usuário
- Habilitar ou desabilitar o monitoramento detalhado da instância
- Monitoramento aprimorado
- Monitoramento de grupos e instâncias do Auto Scaling usando o Amazon CloudWatch
- Publicar métricas personalizadas

- Uso dos alarmes do Amazon CloudWatch
- Uso de painéis do CloudWatch
- Uso de painéis do CloudWatch entre regiões e contas
- Uso do rastreamento do X-Ray entre regiões e contas
- Compreensão da disponibilidade
- Implementação do Amazon Health Aware (AHA)

Vídeos relacionados:

- Mitigating gray failures (Mitigando falhas cinzentas)

Exemplos relacionados:

- Laboratório do Well-Architected: nível 300: implementação de verificações de integridade e do gerenciamento de dependências para melhorar a confiabilidade
- Um workshop de observabilidade: explore o X-Ray

Ferramentas relacionadas:

- CloudWatch
- CloudWatch X-Ray

REL11-BP02 Failover para recursos íntegros

Se ocorrer uma falha no recurso, os recursos íntegros deverão continuar atendendo às solicitações. Para falhas de localização (como zona de disponibilidade ou Região da AWS), garanta que você tenha sistemas implementados para realizar failover para recursos íntegros em locais sem problemas.

Ao projetar um serviço, distribua a carga entre recursos, zonas de disponibilidade ou regiões. Portanto, a falha ou a deficiência de um recurso individual podem ser atenuadas transferindo o tráfego para os recursos íntegros restantes. Pense em como os serviços são descobertos e encaminhados em caso de falha.

Projete seus serviços pensando na recuperação de falhas. Na AWS, projetamos os serviços para minimizar o tempo para recuperação de falhas e o impacto sobre os dados. Nossos serviços usam

principalmente datastores que reconhecem solicitações apenas após serem armazenadas de modo durável entre várias réplicas em uma região. 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.

Os padrões e os designs que permitem o failover variam para cada serviço de plataforma da AWS. Muitos serviços gerenciados nativos da AWS são nativamente várias zonas de disponibilidade (como o Lambda ou o API Gateway). Outros serviços da AWS (como EC2 e EKS) exigem designs específicos de práticas recomendadas para oferecer compatibilidade com o failover de recursos ou armazenamento de dados entre AZs.

O monitoramento deve ser configurado para conferir se o recurso de failover está íntegro, acompanhar o andamento do failover dos recursos e monitorar a recuperação do processo empresarial.

Resultado desejado: Os sistemas são capazes de usar novos recursos de forma automática ou manual para se recuperarem da degradação.

Antipadrões comuns:

- Planejar o fracasso não faz parte da fase de planejamento e design.
- O RTO e o RPO não são estabelecidos.
- Monitoramento insuficiente para detectar falhas nos recursos.
- Isolamento adequado dos domínios de falha.
- O failover multirregional não é considerado.
- A detecção de falhas é sensível ou agressiva demais ao decidir realizar o failover.
- Não testar nem validar o design de failover.
- Executar a automação de autorreparação sem notificar que a reparação era necessária.
- Falta de um período de amortecimento a fim de evitar falhas muito precoces.

Benefícios de estabelecer esta prática recomendada: Você pode criar sistemas mais resilientes que mantenham a confiabilidade em caso de falhas, degradando-se normalmente e se recuperando com rapidez.

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

Orientação para implementação

Os serviços da AWS, como o Elastic Load Balancing e o Amazon EC2 Auto Scaling, ajudam a distribuir a 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 workloads multirregionais, os projetos são mais complicados. Por exemplo, réplicas de leitura entre regiões permitem que você implante os dados em várias Regiões da AWS. No entanto, o failover ainda é necessário para promover a réplica de leitura como primária e direcionar seu tráfego para o novo endpoint. O Amazon Route 53, o Route 53, o Route 53 ARC, o CloudFront e o AWS Global Accelerator podem ajudar a direcionar o tráfego entre Regiões da AWS.

Serviços da AWS, como Amazon S3, Lambda, API Gateway, Amazon SQS, Amazon SNS, Amazon SES, Amazon Pinpoint, Amazon ECR, AWS Certificate Manager, EventBridge ou Amazon DynamoDB, são implantados automaticamente em várias zonas de disponibilidade pela AWS. Em caso de falha, esses serviços da AWS direcionam 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, o Amazon Aurora, o Amazon Redshift, o Amazon EKS ou o Amazon ECS, Multi-AZ é uma opção de configuração. A AWS poderá direcionar o tráfego para a instância íntegra se o failover for iniciado. Essa ação de failover pode ser realizada pela AWS ou conforme exigido pelo cliente.

Para instâncias do Amazon EC2, o Amazon Redshift, tarefas do Amazon ECS ou pods do Amazon EKS, você escolhe em quais zonas de disponibilidade implantar. Para alguns designs, o Elastic Load Balancing fornece a solução para detectar instâncias em zonas não íntegras e direcionar o tráfego para as zonas íntegras. O Elastic Load Balancing também pode rotear o tráfego para componentes em seu datacenter on-premises.

Para o failover de tráfego em várias regiões, o redirecionamento pode utilizar o Amazon Route 53, o Route 53 ARC, o AWS Global Accelerator, o Route 53 Private DNS for VPCs ou o CloudFront para oferecer uma maneira de definir domínios da Internet e atribuir políticas de roteamento, incluindo verificações de integridade, para rotear o tráfego para regiões íntegras. O AWS Global Accelerator fornece endereços IP estáticos que atuam como um ponto de entrada fixo para sua aplicação e, depois, são roteados para os endpoints nas Regiões da AWS de sua escolha, usando a rede global da AWS em vez da Internet com o objetivo de melhorar o desempenho e a confiabilidade.

Etapas da implementação

- Crie designs de failover para todas as aplicações e serviços apropriados. Isole cada componente da arquitetura e crie designs de failover que atendam ao RTO e ao RPO de cada componente.
- Configure ambientes inferiores (como desenvolvimento ou teste) com todos os serviços necessários para ter um plano de failover. Implemente as soluções usando a infraestrutura como código (IaC) para garantir a repetibilidade.
- Configure um local de recuperação, como uma segunda região, para implementar e testar os designs de failover. Se necessário, os recursos para testes podem ser configurados temporariamente para limitar os custos adicionais.
- Determine quais planos de failover são automatizados pela AWS, quais podem ser automatizados por um processo de DevOps e quais podem ser manuais. Documente e avalie o RTO e o RPO de cada serviço.
- Crie um manual de failover e inclua todas as etapas para realizar o failover de cada recurso, aplicação e serviço.
- Crie um manual de failback e inclua todas as etapas de failback (com tempo) de cada recurso, aplicação e serviço.
- Crie um plano para iniciar e ensaiar o manual. Use simulações e testes de caos para testar a automação e as etapas do manual.
- Para danos na localização (como zona de disponibilidade ou Região da AWS), garanta que você tenha sistemas implementados para realizar failover para recursos íntegros em locais sem problemas. Confira a cota, os níveis de ajuste de escala automático e os recursos em execução antes do teste de failover.

Recursos

Práticas recomendadas relacionadas ao Well-Architected:

- REL13 – Plano para DR
- REL10 – Usar o isolamento de falhas para proteger a workload

Documentos relacionados:

- Setting RTO and RPO Targets
- Set up Route 53 ARC with application loadbalancers

- Failover using Route 53 Weighted routing
- DR with Route 53 ARC
- EC2 with autoscaling
- EC2 Deployments - Multi-AZ
- ECS Deployments - Multi-AZ
- Switch traffic using Route 53 ARC
- Lambda with an Application Load Balancer and Failover
- ACM Replication and Failover
- Parameter Store Replication and Failover
- ECR cross region replication and Failover
- Secrets manager cross region replication configuration
- Enable cross region replication for EFS and Failover
- EFS Cross Region Replication and Failover
- Networking Failover
- S3 Endpoint failover using MRAP
- Create cross region replication for S3
- Failover Regional API Gateway with Route 53 ARC
- Failover using multi-region global accelerator
- Failover with DRS
- Creating Disaster Recovery Mechanisms Using Amazon Route 53

Exemplos relacionados:

- Recuperação de desastres na AWS
- Recuperação elástica de desastres na AWS

REL11-BP03 Automatizar a reparação em todas as camadas

Após a detecção de uma falha, use recursos automatizados para executar ações de correção. As degradações podem ser corrigidas automaticamente por meio de mecanismos internos de serviço ou exigir que os recursos sejam reiniciados ou removidos por meio de ações de remediação.

Para aplicações autogerenciadas e reparação entre regiões, os projetos de recuperação e os processos de recuperação automatizados podem ser extraídos de práticas recomendadas existentes.

A capacidade de reiniciar ou remover um recurso é uma ferramenta importante para corrigir falhas. Uma prática recomendada é deixar os serviços sem estado sempre que possível. Isso evita a perda de dados ou a disponibilidade na reinicialização do recurso. Na nuvem, você pode (e geralmente deve) substituir todo o recurso (por exemplo, uma instância de computação ou função sem servidor) 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.

Reiniciar ou tentar novamente também se aplica às solicitações 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. A capacidade de reiniciar é um mecanismo de recuperação presente na computação orientada para a recuperação e arquiteturas de cluster de alta disponibilidade.

Resultado desejado: Ações automatizadas são executadas para corrigir a detecção de uma falha.

Antipadrões comuns:

- Provisionamento de recursos sem dimensionamento automático.
- 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 do ajuste de escala automático e recuperação automática.
- Sem automação para failover nos bancos de dados.
- Não há métodos automatizados para redirecionar o tráfego para novos endpoints.
- Sem replicação de armazenamento.

Benefícios de estabelecer esta prática recomendada: A reparação automatizada pode reduzir seu tempo médio de recuperação e melhorar sua disponibilidade.

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

Orientação para implementação

Os designs para Amazon EKS ou outros serviços do Kubernetes devem incluir conjuntos mínimos e máximos de réplicas ou com estado e tamanho mínimo do cluster e do grupo de nós. Esses mecanismos fornecem uma quantidade mínima de recursos de processamento continuamente disponíveis e, ao mesmo tempo, remediam automaticamente quaisquer falhas usando o ambiente de gerenciamento do Kubernetes.

Os padrões de design que são acessados por meio de um balanceador de carga usando clusters de computação devem utilizar grupos Auto Scaling. O Elastic Load Balancing (ELB) distribui automaticamente o tráfego de entrada da aplicação entre vários destinos e dispositivos virtuais em uma ou mais Zonas de Disponibilidade (AZs).

Designs baseados em computação em cluster que não usam balanceamento de carga devem ter seu tamanho projetado para a perda de pelo menos um nó. Isso permitirá que o serviço se mantenha funcionando em uma capacidade potencialmente reduzida enquanto recupera um novo nó. Entre os exemplos de serviço estão Mongo, DynamoDB Accelerator, Amazon Redshift, Amazon EMR, Cassandra, Kafka, MSK-EC2, Couchbase, ELK e Amazon OpenSearch Service. Muitos desses serviços podem ser projetados com recursos adicionais de recuperação automática. Algumas tecnologias de cluster devem gerar um alerta sobre a perda de um nó, acionando um fluxo de trabalho automático ou manual para recriar um novo nó. Esse fluxo de trabalho pode ser automatizado usando o AWS Systems Manager para corrigir problemas rapidamente.

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 então invocar AWS Lambda, Systems Manager Automation ou outros destinos para executar uma lógica de remediação personalizada na workload. O Amazon EC2 Auto Scaling pode ser configurado para verificar a integridade da instância 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. Para substituições em grande escala (como a perda de uma Zona de Disponibilidade inteira), a estabilidade estática é preferida para alta disponibilidade.

Etapas da implementação

- Use grupos do Auto Scaling para implantar camadas em uma workload. O Auto Scaling pode executar a autorreparação em aplicações sem estado e adicionar e remover capacidade.

- Para instâncias computacionais mencionadas anteriormente, use balanceamento de carga e escolha o tipo apropriado de balanceador de carga.
- Considere a recuperação para Amazon RDS. Com instâncias em espera, configure para failover automático para a instância em espera. Para a réplica de leitura do Amazon RDS, é necessário um fluxo de trabalho automatizado para tornar uma réplica de leitura primária.
- Implemente recuperação automática em instâncias do EC2 que têm aplicações implantadas que não podem estar em vários locais e podem tolerar a reinicialização em caso de falhas. É possível usar a recuperação automática para substituir o hardware com falha e reiniciar a instância quando a aplicação não puder ser implantada em vários locais. Os metadados da instância e os endereços IP associados são mantidos, bem como os volumes do EBS e pontos de montagem no Amazon Elastic File System ou sistemas de arquivos para Lustre e Windows. Com o uso do AWS OpsWorks, é possível configurar a autorreparação das instâncias do EC2 no nível da camada.
- Implemente a recuperação automatizada usando o AWS Step Functions e o AWS Lambda quando não for possível usar o ajuste de escala automático ou a recuperação automática, ou quando a recuperação automática falhar. Quando não for possível usar o ajuste de escala automático e a recuperação automática ou quando a recuperação automática falhar, você poderá automatizar a reparação usando o AWS Step Functions e o AWS Lambda.
- O Amazon EventBridge pode ser usado para monitorar e filtrar eventos como alarmes do CloudWatch ou mudanças de estado em outros serviços da AWS. Com base nas informações do evento, ele pode invocar o AWS Lambda (ou outros destinos) para executar a lógica de correção personalizada na workload.

Recursos

Práticas recomendadas relacionadas:

- Definição de disponibilidade
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

Documentos relacionados:

- Como funciona o AWS Auto Scaling
- 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?
- AWS OpsWorks: como usar a correção automática para substituir instâncias com falha
- O que é o AWS Step Functions?
- O que é o AWS Lambda?
- O que é o Amazon EventBridge?
- Uso dos alarmes do Amazon CloudWatch
- Failover do Amazon RDS
- SSM - Automação do Systems Manager
- Práticas recomendadas de arquitetura resiliente

Vídeos relacionados:

- Provisionar e escalar automaticamente o OpenSearch Service
- Failover automático do Amazon RDS

Exemplos relacionados:

- Workshop sobre Auto Scaling
- Workshop de failover do Amazon RDS

Ferramentas relacionadas:

- CloudWatch
- CloudWatch X-Ray

REL11-BP04 Confiar no plano de dados e não no ambiente de gerenciamento durante a recuperação

Os ambientes de gerenciamento fornecem as APIs administrativas usadas para criar, ler e descrever, atualizar, excluir e listar recursos (CRUDL), enquanto os planos de dados lidam com o tráfego diário de serviços. Ao implementar respostas de recuperação ou mitigação a eventos potencialmente impactantes na resiliência, concentre-se em usar um número mínimo de operações do ambiente de

gerenciamento para recuperar, redimensionar, restaurar, reparar ou realizar o failover do serviço. A ação do plano de dados deve substituir qualquer atividade durante esses eventos de degradação.

Por exemplo, estas são ações do ambiente de gerenciamento: iniciar uma nova instância de computação, criar armazenamento em bloco e descrever serviços de fila. Quando você executa instâncias de computação, o ambiente de gerenciamento precisa realizar várias tarefas, como encontrar um host físico com capacidade, alocar interfaces de rede, preparar volumes de armazenamento em blocos locais, gerar credenciais e adicionar regras de segurança. Os ambientes de gerenciamento tendem a ser uma orquestração complicada.

Resultado desejado: Quando um recurso entra em um estado comprometido, o sistema é capaz de se recuperar automática ou manualmente, transferindo o tráfego de recursos danificados para recursos saudáveis.

Antipadrões comuns:

- Dependência da alteração dos registros DNS para redirecionar o tráfego.
- Dependência das operações de escalonamento do ambiente de gerenciamento para substituir componentes danificados devido a recursos insuficientemente provisionados.
- Dependência de ações de ambiente de gerenciamento abrangentes, com vários serviços e várias APIs para remediar qualquer categoria de deficiência.

Benefícios de estabelecer esta prática recomendada: O aumento da taxa de sucesso da remediação automatizada pode reduzir seu tempo médio de recuperação e melhorar a disponibilidade da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: médio: para certos tipos de degradação do serviço, os ambientes de gerenciamento são afetados. As dependências do uso extensivo do ambiente de gerenciamento para remediação podem aumentar o tempo de recuperação (RTO) e o tempo médio de recuperação (MTTR).

Orientação para implementação

Para limitar as ações do plano de dados, avalie cada serviço quanto às ações necessárias para restaurar o serviço.

Use o Amazon Route 53 Application Recovery Controller para mudar o tráfego de DNS. 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.

As políticas de roteamento do Route 53 usam o ambiente de gerenciamento. Portanto, não confie nele para 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: Leste dos EUA (Norte da Virgínia). 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 o Amazon EC2, use designs de estabilidade estática para limitar as ações do ambiente de gerenciamento. As ações do ambiente de gerenciamento incluem a ampliação de recursos individualmente ou usando grupos do Auto Scaling (ASG). Para obter os níveis mais altos de resiliência, provisione capacidade suficiente no cluster usado para failover. Se essa capacidade precisar ser limitada, defina valores no sistema geral completo para definir com segurança o tráfego total que atinge o conjunto limitado de recursos.

Para serviços como Amazon DynamoDB, Amazon API Gateway, balanceadores de carga e AWS Lambda sem servidor, o uso aproveita o plano de dados. No entanto, criar novas funções, balanceadores de carga, gateways de API ou tabelas de DynamoDB é uma ação do ambiente de gerenciamento e deve ser concluída antes da degradação, como preparação para um evento e ensaio das ações de failover. Para o Amazon RDS, as ações do plano de dados permitem o acesso aos dados.

Para obter mais informações sobre planos de dados, ambientes de gerenciamento e como a AWS cria serviços para atender metas de alta disponibilidade, consulte Estabilidade estática usando Zonas de disponibilidade.

Entenda quais operações estão no plano de dados e quais estão no ambiente de gerenciamento.

Etapas da implementação

Para cada workload que precisa ser restaurada após um evento de degradação, avalie o runbook de failover, o projeto de alta disponibilidade, o projeto de recuperação automática ou o plano de restauração de recursos de HA. Identifique cada ação que pode ser considerada uma ação do ambiente de gerenciamento.

Considere alterar a ação de gerenciamento para uma ação do plano de dados:

- Auto Scaling (ambiente de gerenciamento) em comparação com recursos do Amazon EC2 pré-escalados (plano de dados)
- Migre para Lambda e seus métodos de escalabilidade (plano de dados) ou Amazon EC2 e ASG (ambiente de gerenciamento)
- Avalie qualquer projeto usando o Kubernetes e a natureza das ações do ambiente de gerenciamento. Adicionar pods é uma ação do plano de dados no Kubernetes. As ações devem se limitar à adição de pods e não adição de nós. O uso de nós com provisionamento excessivo é o método preferido para limitar as ações do ambiente de gerenciamento

Considere abordagens alternativas que permitam que as ações do plano de dados afetem a mesma remediação.

- Alteração de registro Route 53 (ambiente de gerenciamento) ou Route 53 ARC (plano de dados)
- Verificações de integridade do Route 53 para atualizações mais automatizadas

Se o serviço for essencial, considere alguns serviços em uma região secundária para permitir mais ações no ambiente de gerenciamento e no plano de dados em uma região não afetada.

- Amazon EC2 Auto Scaling ou Amazon EKS em uma região primária em comparação com Amazon EC2 Auto Scaling ou Amazon EKS em uma região secundária e roteando o tráfego para a região secundária (ação do ambiente de gerenciamento).
- Faça uma réplica de leitura na primária secundária ou tente a mesma ação na região primária (ação do ambiente de gerenciamento).

Recursos

Práticas recomendadas relacionadas:

- Definição de disponibilidade
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas

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?
- Ambiente de gerenciamento e plano de dados do Kubernetes

Vídeos relacionados:

- Back to Basics - Using Static Stability (De volta ao básico: uso da estabilidade estática)
- Building resilient multi-site workloads using AWS global services (Criação de workloads resilientes em vários sites usando serviços globais da AWS)

Exemplos relacionados:

- Introdução ao Amazon Route 53 Application Recovery Controller
- Amazon Builders' Library: evite a sobrecarga em sistemas distribuídos colocando o menor serviço no controle
- 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
- Estabilidade estática usando Zonas de disponibilidade

Ferramentas relacionadas:

- Amazon CloudWatch
- AWS X-Ray

REL11-BP05 Usar estabilidade estática para evitar o comportamento bimodal

As workloads devem ser estaticamente estáveis e operar somente em um único modo normal. O comportamento bimodal ocorre quando a workload exibe um comportamento diferente nos modos normal e de falha.

Por exemplo, você pode tentar se recuperar de uma falha na zona de disponibilidade iniciando novas instâncias em uma zona de disponibilidade diferente. Isso pode resultar em uma resposta bimodal durante um modo de falha. Em vez disso, você deve criar workloads que sejam estaticamente estáveis e que operem em apenas um modo. Neste exemplo, essas instâncias deveriam ter sido provisionadas na segunda zona de disponibilidade antes da falha. Esse design de estabilidade estática verifica se a workload opera somente em um único modo.

Resultado desejado: As workloads não apresentam comportamento bimodal durante os modos normal e de falha.

Antipadrões comuns:

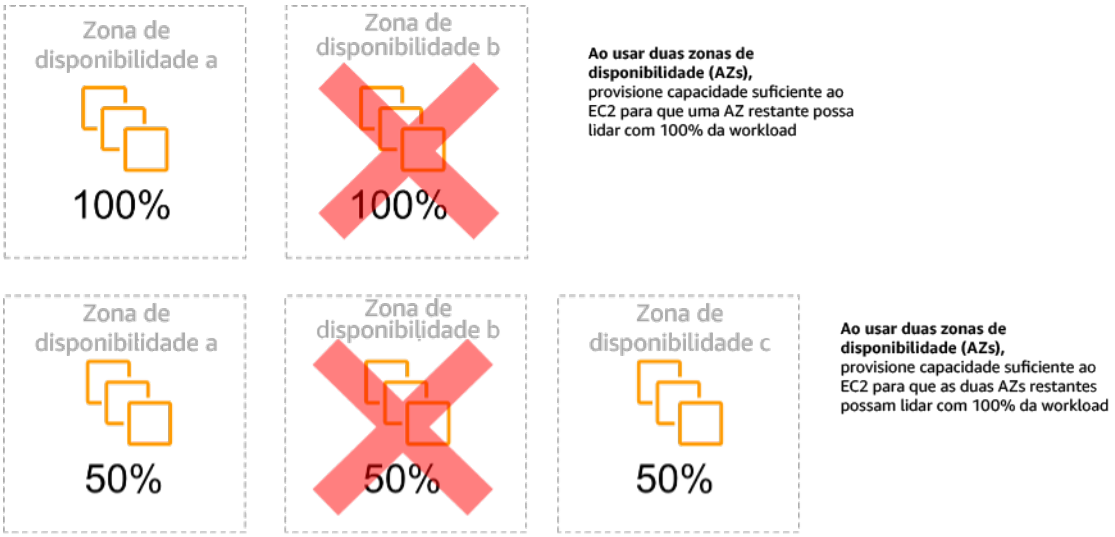
- Supor que os recursos sempre possam ser provisionados, independentemente do escopo da falha.
- Tentar adquirir recursos dinamicamente durante uma falha.
- Não provisionar recursos adequados entre zonas ou regiões até que ocorra uma falha.
- Pensar em projetos estáticos estáveis somente para recursos computacionais.

Benefícios de estabelecer esta prática recomendada: As workloads executadas com projetos estaticamente estáveis podem ter resultados previsíveis durante eventos normais e de falha.

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

Orientação para implementação

O comportamento bimodal ocorre quando a workload apresenta um comportamento diferente nos modos normal e de falha (por exemplo, depender da inicialização de novas instâncias se uma zona de disponibilidade falhar). Um exemplo de comportamento bimodal é quando projetos estáveis do Amazon EC2 provisionam instâncias suficientes em cada zona de disponibilidade para lidar com a carga da workload se uma AZ for removida. A integridade do Elastic Load Balancing ou do Amazon Route 53 conferiria a possibilidade de isolar a carga das instâncias prejudicadas. 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. A estabilidade estática para implantação de computação (como instâncias ou contêineres do EC2) resulta na mais alta confiabilidade.



Estabilidade estática de instâncias do EC2 em várias zonas de disponibilidade

Isso deve ser comparado ao custo desse modelo e ao valor comercial de manter a workload em todos os casos de resiliência. É mais barato provisionar menos capacidade computacional e depender da inicialização de novas instâncias em caso de falha. No entanto, para falhas em grande escala (como um dano regional ou da zona de disponibilidade), essa abordagem é menos eficaz porque depende tanto de um plano operacional quanto de recursos suficientes disponíveis nas zonas ou regiões não afetadas.

A solução também deve comparar a confiabilidade com os custos necessários para a workload. As arquiteturas de estabilidade estática se aplicam a uma variedade de arquiteturas, incluindo instâncias de computação espalhadas por zonas de disponibilidade, projetos de réplicas de leitura de banco de dados, projetos de cluster do Kubernetes (Amazon EKS) e arquiteturas de failover de várias regiões.

Também é possível implementar um design mais estável estaticamente usando mais recursos em cada zona. Ao adicionar mais zonas, você reduz a quantidade de computação adicional necessária para a estabilidade estática.

Um exemplo de comportamento bimodal seria um tempo limite de rede que poderia fazer com que um sistema tentasse atualizar seu próprio estado de configuração por completo. Isso adicionaria uma carga inesperada a outro componente e poderia fazê-lo falhar, resultando em outras consequências inesperadas. Esse ciclo de feedback negativo afeta a disponibilidade da workload. Em vez disso, você pode criar sistemas estaticamente estáveis e operar em apenas um modo. Um design estático estável faria um trabalho constante e sempre atualizaria o estado da configuração em um ritmo fixo. Quando uma chamada falha, a workload usa o valor previamente armazenado em cache e inicia um alarme.

Outro exemplo de comportamento bimodal é permitir que os clientes ignorem o cache da workload em caso de falhas. Essa pode parecer uma solução que acomoda as necessidades do cliente, mas pode alterar significativamente as demandas da workload e provavelmente resultar em falhas.

Avalie workloads importantes para determinar quais workloads exigem esse tipo de projeto de resiliência. Para as que são consideradas críticas, cada componente da aplicação deve ser revisado. Exemplos de tipos de serviço que exigem avaliações de estabilidade estática são:

- Computação: Amazon EC2, EKS-EC2, ECS-EC2, EMR-EC2
- Bancos de dados: Amazon Redshift, Amazon RDS, Amazon Aurora
- Armazenamento: Amazon S3 (Zona única), Amazon EFS (montagens), Amazon FSx (montagens)
- Balanceadores de carga: Em determinados projetos

Etapas da implementação

- Crie sistemas que sejam estaticamente estáveis e que operem em apenas um modo. Nesse caso, provisione instâncias suficientes em cada zona de disponibilidade ou região para lidar com a capacidade da workload se uma zona de disponibilidade ou região for removida. Uma variedade de serviços pode ser usada para roteamento a recursos íntegros, como:
  - Roteamento de DNS entre regiões
  - Roteamento de várias regiões do Amazon S3
  - AWS Global Accelerator
  - Amazon Route 53 Application Recovery Controller

- Configure réplicas de leitura do banco de dados para contabilizar a perda de uma única instância principal ou de uma réplica de leitura. Se o tráfego estiver sendo servido por réplicas de leitura, a quantidade em cada zona de disponibilidade e cada região deve ser igual à necessidade geral em caso de falha na zona ou região.
- Configure dados importantes no armazenamento do Amazon S3, projetado para ser estaticamente estável para dados armazenados em caso de falha na zona de disponibilidade. Se Se a classe de armazenamento Amazon S3 One Zone-IA for usada, ela não deverá ser considerada estaticamente estável, pois a perda dessa zona minimiza o acesso a esses dados armazenados.
- Às vezes, os balanceadores de carga são configurados incorretamente ou intencionalmente para atender a uma zona de disponibilidade específica. Nesse caso, o design estaticamente estável pode envolver a distribuição de uma workload entre várias zonas de disponibilidade em um design mais complexo. O design original pode ser usado para reduzir o tráfego entre zonas por motivos de segurança, latência ou custo.

Recursos

Práticas recomendadas relacionadas ao Well-Architected:

- Definição de disponibilidade
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP04 Confiar no plano de dados e não no ambiente de gerenciamento durante a recuperação

Documentos relacionados:

- Minimizar dependências em um plano de recuperação de desastres
- A Amazon Builders' Library: estabilidade estática usando zonas de disponibilidade
- Limites de isolamento de falhas
- Estabilidade estática usando Zonas de disponibilidade
- Multi-AZ do Amazon RDS
- Minimizar dependências em um plano de recuperação de desastres
- Roteamento de DNS entre regiões
- Roteamento de várias regiões do Amazon S3
- AWS Global Accelerator

- Route 53 ARC
- Zona única do Amazon S3
- Balanceamento de carga entre zonas

Vídeos relacionados:

- Static stability in AWS: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)

Exemplos relacionados:

- A Amazon Builders’ Library: estabilidade estática usando zonas de disponibilidade

REL11-BP06 Enviar notificações quando os eventos afetarem a disponibilidade

As notificações são enviadas após a detecção de limites violados, mesmo que o evento causado pelo problema tenha sido resolvido automaticamente.

A correção automatizada permite que a workload 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.

Os sistemas resilientes são projetados para que os eventos de degradação sejam comunicados imediatamente às equipes apropriadas. Essas notificações devem ser enviadas por meio de um ou vários canais de comunicação.

Resultado desejado: Os alertas são enviados imediatamente às equipes de operações quando os limites são violados. Esses alertas podem incluir taxas de erro, latência ou outras métricas importantes de indicadores-chave de performance (KPI), permitindo que esses problemas sejam resolvidos o mais rápido possível e o impacto do usuário seja evitado ou minimizado.

Antipadrões comuns:

- Enviar muitos alarmes.
- Enviar alarmes que não resultam em ações concretas.
- Definir limites de alarme muito altos (supersensíveis) ou muito baixos (subsensíveis).
- Não enviar alarmes para dependências externas.

- Não considerar falhas cinzentas ao projetar monitoramento e alarmes.
- Executar a automação da correção, mas sem notificar a equipe apropriada de que a correção era necessária.

Benefícios de estabelecer esta prática recomendada: As notificações de recuperação alertam as equipes operacionais e empresariais sobre as degradações do serviço, para que possam reagir imediatamente a fim de minimizar o tempo médio de detecção (MTTD) e o tempo médio de reparo (MTTR). As notificações de eventos de recuperação também garantem que você não ignore problemas que ocorrem com pouca frequência.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio. A falha na implementação de mecanismos adequados de monitoramento e notificação de eventos pode resultar em falha na detecção de padrões de problemas, incluindo aqueles resolvidos pela recuperação automática. Uma equipe só será informada da degradação do sistema quando os usuários entrarem em contato com o atendimento ao cliente ou por acaso.

Orientações para a implementação

Ao definir uma estratégia de monitoramento, um alarme acionado é um evento comum. Esse evento provavelmente conteria um identificador para o alarme, o estado do alarme (como `EM ALARME` ou `OK`) e detalhes do que o desencadeou. Em muitos casos, deve-se detectar um evento de alarme e enviar uma notificação por e-mail. Este é um exemplo de uma ação em um alarme. A notificação do alarme é fundamental para a observabilidade, pois informa às pessoas certas que há um problema. No entanto, quando a ação sobre eventos amadurece em sua solução de observabilidade, ela pode corrigir automaticamente o problema sem a necessidade de intervenção humana.

Depois que os alarmes de monitoramento de KPI forem estabelecidos, os alertas deverão ser enviados às equipes apropriadas quando os limites forem excedidos. Esses alertas também podem ser usados para acionar processos automatizados que tentarão remediar a degradação.

Para um monitoramento de limite mais complexo, considere alarmes compostos. Eles usam vários alarmes de monitoramento de KPI para criar um alerta com base na lógica operacional de negócios. Os 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 ou Amazon EventBridge.

Etapas para a implementação

Crie vários tipos de alarme com base na forma como as workloads são monitoradas, por exemplo:

- Os alarmes de aplicações são usados para detectar quando alguma parte da workload não está funcionando adequadamente.
- Alarmes de infraestrutura indicam quando escalar os recursos. Os alarmes podem ser exibidos visualmente em painéis, enviar alertas pelo Amazon SNS ou por e-mail e trabalhar com o Auto Scaling para aumentar ou reduzir a escala dos recursos da workload horizontalmente.
- Simples alarmes estáticos podem ser criados para monitorar quando uma métrica ultrapassa um limite estático durante um número específico de períodos de avaliação.
- Os alarmes compostos podem representar alarmes complexos de várias origens.
- Depois que o alarme for criado, crie eventos de notificação apropriados. Você pode invocar diretamente uma API do Amazon SNS para enviar notificações e vincular qualquer automação para correção ou comunicação.
- integre o Amazon Health Aware para permitir a visibilidade de monitoramento de recursos da AWS que possam ter degradações. Para workloads essenciais aos negócios, essa solução fornece acesso a alertas proativos e em tempo real para serviços da AWS.

Recursos

Práticas recomendadas relacionadas ao Well-Architected:

- Definição de disponibilidade

Documentos relacionados:

- Criar um alarme do CloudWatch com base em um limite estático
- O que é o Amazon EventBridge?
- O que é o Amazon Simple Notification Service?
- Publicar métricas personalizadas
- Uso dos alarmes do Amazon CloudWatch
- Amazon Health Aware (AHA)
- Setup CloudWatch Composite alarms
- What's new in AWS Observability at re:Invent 2022

Ferramentas relacionadas:

- CloudWatch
- CloudWatch X-Ray

REL11-BP07 Arquitetar o produto para cumprir as metas de disponibilidade e os acordos de nível de serviço (SLAs) de tempo de atividade

Arquitete o produto para cumprir as metas de disponibilidade e os acordos de nível de serviço (SLAs) de tempo de atividade. Se você publicar ou concordar de forma privada com as metas de disponibilidade ou SLAs de tempo de atividade, verifique se sua arquitetura e seus processos operacionais foram projetados para comportá-los.

Resultado desejado: cada aplicação tem uma meta definida com relação à disponibilidade e SLA para métricas de desempenho, o que pode ser monitorado e mantido para cumprir os resultados empresariais.

Antipadrões comuns:

- Planejar e implantar workloads sem definir SLAs.
- As métricas de SLA são definidas como altas sem justificativa ou requisitos empresariais.
- Definir SLAs sem considerar as dependências e o SLA subjacente.
- Os designs da aplicação são criados sem considerar o modelo de responsabilidade compartilhada para resiliência.

Benefícios do estabelecimento dessa prática recomendada: projetar aplicações com base nas principais metas de resiliência ajuda a cumprir os objetivos empresariais e atender às expectativas dos clientes. Esses objetivos ajudam a orientar o processo de design da aplicação que avalia diferentes tecnologias e considera as vantagens e desvantagens.

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

Orientação de implementação

Os designs da aplicação precisam levar em conta um conjunto de requisitos diversos que são derivados dos objetivos empresariais, operacionais e financeiros. Nos requisitos operacionais, as workloads precisam ter metas de métricas de resiliência específicas para que possam ser monitorados e comportados adequadamente. As métricas de resiliência não devem ser definidas nem derivadas depois de implantar a workload. Elas devem ser definidas durante a fase de design e ajudar a orientar as diversas decisões e concessões.

- Cada workload deve ter seu próprio conjunto de métricas de resiliência. Essas métricas podem ser diferentes de outras aplicações empresariais.
- Reduzir as dependências pode ter um impacto positivo na disponibilidade. Cada workload deve considerar suas dependências e seus SLAs. Em geral, escolha dependências com metas de disponibilidade iguais ou maiores que as metas da workload.
- Considere designs com acoplamento fraco para que a workload possa operar corretamente apesar do comprometimento da dependência, quando possível.
- Reduza as dependências do ambiente de gerenciamento, especialmente durante uma recuperação ou degradação. Avalie os designs estaticamente estáveis com relação às workloads essenciais à missão. Use a economia de recursos para aumentar a disponibilidade dessas dependências em uma workload.
- A capacidade de observação e a instrumentalização são críticas para cumprir os SLAs reduzindo o tempo médio de detecção (MTTD) e o tempo médio de reparo (MTTR).
- Falha menos frequente (MTBF mais longo), tempo de detecção de falhas mais curto (MTTD mais curto) e tempo de reparo mais curto (MTTR mais curto) são os três fatores usados para melhorar a disponibilidade em sistemas distribuídos.
- Estabelecer e cumprir métricas de resiliência para uma workload é fundamental para qualquer design eficaz. Esses designs devem levar em consideração as vantagens e desvantagens da complexidade de design, as dependências do serviço, o desempenho, a escalabilidade e os custos.

Etapas da implementação

- Analise e documente o design da workload considerando as seguintes questões:
  - Onde são usados ambientes de gerenciamento na workload?
  - Como a workload implementa tolerância a falhas?
  - Quais são os padrões de design para componentes de escalabilidade, escalabilidade automática, redundância e alta disponibilidade?
  - Quais são os requisitos para disponibilidade e consistência de dados?
  - Há considerações quanto à economia de recursos ou estabilidade estática de recursos?
  - Quais são as dependências do serviço?
- Defina métricas de SLA com base na arquitetura da workload enquanto trabalha com as partes interessadas. Considere os SLAs de todas as dependências usadas pela workload.
- Quando a meta de SLA for definida, otimize a arquitetura para cumprir o SLA.

- Quando for definido o design que cumprirá o SLA, implemente mudanças operacionais, automação do processo e runbooks que também terão como foco uma redução de MTTD e MTTR.
- Depois da implantação, monitore e informe sobre o SLA.

Recursos

Práticas recomendadas relacionadas:

- REL03-BP01 Escolher como segmentar a workload
- REL10-BP01 Implantar a workload em vários locais
- REL11-BP01 Monitorar todos os componentes da workload para detectar falhas
- REL11-BP03 Automatizar a reparação em todas as camadas
- REL12-BP05 Testar a resiliência por meio da engenharia do caos
- REL13-BP01 Definir os objetivos de recuperação para tempo de inatividade e perda de dados
- Compreensão de integridade da workload

Documentos relacionados:

- Availability with redundancy (Disponibilidade com redundância)
- Pilar Confiabilidade: disponibilidade
- Measuring availability (Medição da disponibilidade)
- AWS Fault Isolation Boundaries (Limites de isolamento de falhas da AWS)
- Modelo de responsabilidade compartilhada para resiliência
- estabilidade estática usando Zonas de disponibilidade
- AWS Service Level Agreements (SLAs) (Acordos de Nível de Serviço (SLAs) da AWS)
- Guidance for Cell-based Architecture on AWS(Orientações para arquitetura baseada em células com a AWS)
- Infraestrutura da AWS
- Advanced Multi-AZ Resiliance Patterns whitepaper (Whitepaper de padrões de resiliência Multi-AZ avançados)

Serviços relacionados:

- Amazon CloudWatch

- AWS Config
- AWS Trusted Advisor

CONFIABILIDADE 12. Como testar a confiabilidade?

Depois de projetar a workload 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.

Resultado desejado: suas equipes têm uma abordagem consistente e acordada para lidar com a análise pós-incidente. Um mecanismo é o processo de correção de erros (COE). O processo de COE ajuda as equipes a identificar, compreender e abordar as causas básicas dos incidentes, ao mesmo tempo em que cria mecanismos e barreiras de proteção para limitar a probabilidade do mesmo incidente ocorrer novamente.

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.

- Concentrar-se em atribuir a culpa em vez de compreender a causa raiz, criando uma cultura de medo e impedindo a comunicação aberta.
- Não compartilhar insights, o que mantém as descobertas da análise de incidentes em um pequeno grupo e impede que outras pessoas se beneficiem das lições aprendidas.
- Não ter um mecanismo para capturar conhecimento institucional e, dessa forma, perder insights valiosos por não preservar as lições aprendidas na forma de práticas recomendadas atualizadas e resultando em incidentes repetidos com a mesma causa raiz ou similar.

Benefícios do estabelecimento dessa prática recomendada: a realização de análises pós-incidentes e o compartilhamento dos resultados permitem que outras workloads atenuem o risco caso tenham implementado os mesmos fatores contribuintes, além de possibilitar que elas implementem a mitigação ou a recuperação automatizada antes que ocorra um incidente.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: alto

Orientações para a implementação

Uma boa análise pós-incidente oferece oportunidades para propor soluções comuns a problemas com padrões de arquitetura usados em outros locais nos sistemas.

A base do processo da COE é documentar e resolver problemas. É recomendável definir uma forma padronizada de documentar as causas raízes essenciais e garantir que elas sejam analisadas e abordadas. Atribua uma propriedade clara ao processo de análise pós-incidente. Designe uma equipe ou uma pessoa responsável para supervisionar as investigações e o acompanhamento de incidentes.

Incentive uma cultura que se concentre no aprendizado e na melhoria, em vez de na atribuição de culpas. Enfatize que a meta é evitar futuros incidentes, não penalizar pessoas.

Desenvolva procedimentos bem definidos para conduzir análises pós-incidentes. Esses procedimentos devem descrever as etapas a serem seguidas, as informações a serem coletadas e as principais questões a serem abordadas durante a análise. Investigue os incidentes minuciosamente, indo além das causas imediatas para identificar as causas raízes e os fatores contribuintes. Use técnicas, como os cinco porquês, para se aprofundar nos problemas subjacentes.

Mantenha um repositório das lições aprendidas com as análises dos incidentes. Esse conhecimento institucional pode servir como referência para futuros incidentes e iniciativas de prevenção. Compartilhe descobertas e insights de análises pós-incidentes e considere realizar reuniões abertas sobre a revisão pós-incidente para discutir as lições aprendidas.

Etapas da implementação

- Ao conduzir a análise pós-incidente, verifique se o processo está livre de culpabilização. Isso permite que as pessoas envolvidas no incidente sejam imparciais com as ações corretivas propostas e promovam uma autoavaliação honesta e a colaboração entre as equipes.
- Defina uma forma padronizada de documentar problemas essenciais. Um exemplo de estrutura para esse documento é o seguinte:
  - O que aconteceu?
  - Qual foi o impacto nos clientes e em sua empresa?
  - Qual foi a causa raiz?
  - Quais dados você tem para apoiar isso?
    - Por exemplo, métricas e grafos
  - Quais foram as implicações críticas nos pilares, especialmente em relação à segurança?
    - Ao arquitetar workloads, você faz concessões entre os pilares com base no contexto da sua empresa. Essas decisões de negócios podem definir suas prioridades de engenharia. Você pode reduzir custos e assim diminuir a confiabilidade em ambientes de desenvolvimento, ou otimizar a confiabilidade e aumentar os custos para soluções importantes. A segurança é sempre prioritária, porque você precisa proteger seus clientes.
  - Quais lições você aprendeu?
  - Quais ações corretivas você está tomando?
    - Itens de ação
    - Itens relacionados
- Crie procedimentos operacionais padrão bem definidos para conduzir análises pós-incidentes.
- Configure um processo padronizado de relatórios de incidentes. Documente todos os incidentes de forma abrangente, incluindo o relatório inicial do incidente, logs, comunicações e ações tomadas durante o incidente.
- Lembre-se de que um incidente não exige uma interrupção. Pode ser uma quase falha ou um sistema que, embora esteja funcionando de forma inesperada, cumpre sua função de negócios.
- Melhore continuamente o processo de análise pós-incidente com base no feedback e nas lições aprendidas.
- Capture as principais descobertas em um sistema de gerenciamento de conhecimento e considere os padrões que devem ser adicionados aos guias de desenvolvedor ou às listas de verificação de pré-implantação.

Recursos

Documentos relacionados:

- Why you should develop a correction of error (COE)

Vídeos relacionados:

- Amazon's approach to failing successfully
- AWS re:Invent 2021 - Amazon Builders' Library: Operational Excellence at Amazon

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.

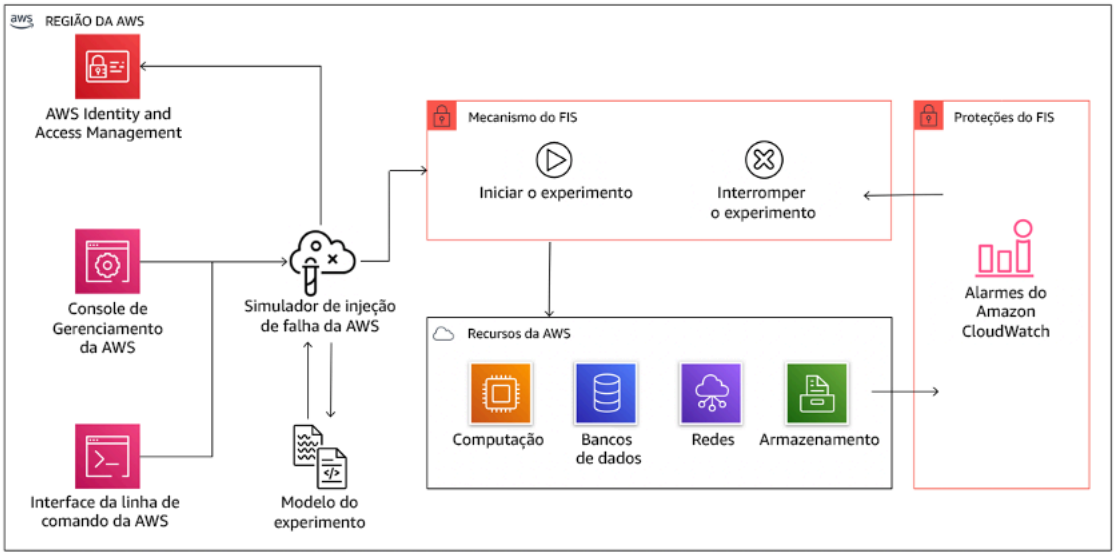
Para falhas no nível da aplicação (como travamentos), você pode começar com fatores de estresse, como exaustão de memória e CPU.

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.

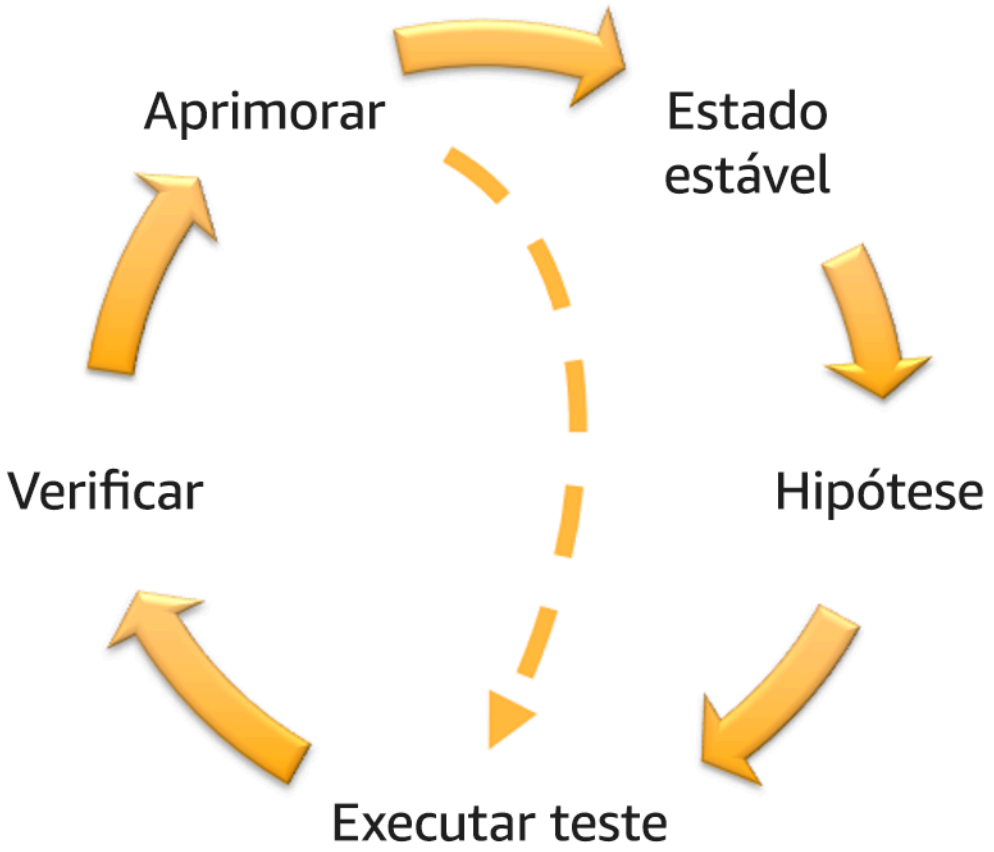
  Ao considerar a frequência de determinada falha, analise os dados passados para essa workload sempre que disponíveis. Caso contrário, use os dados de outras workloads executadas em ambientes semelhantes.

  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 da 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

CONFIABILIDADE 13. Como planejar 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.

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 cumprir os objetivos de recuperação

- REL13-BP03 Testar a implementação da 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 cumprir os 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 cumprir os objetivos de recuperação”
- the section called “REL13-BP03 Testar a implementação da 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 cumprir os objetivos de recuperação

Defina uma estratégia de recuperação de desastres (DR) que cumpra os objetivos de recuperação da workload. Escolha uma estratégia como backup e restauração, standby (ativo/passivo) ou ativo/ativo.

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 ter um plano para a recuperação de desastres.
- Depender das operações do ambiente de gerenciamento durante a recuperação.

Benefícios do estabelecimento desta prática recomendada:

- O uso de estratégias de recuperação definidas permite que você adote ferramentas comuns e procedimentos de teste.
- Usar estratégias de recuperação definidas melhora o compartilhamento de conhecimento entre as equipes e a implementação da DR nas workloads que possuem.

Nível de exposição a riscos quando esta prática recomendada não é estabelecida: alto. Sem uma estratégia de DR planejada, implementada e testada, é improvável que você cumpra os objetivos de recuperação em caso de desastre.

Orientação de implementação

Uma estratégia de DR depende da capacidade de manter a workload em um site de recuperação se o 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 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 estratégias a seguir. 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 cinco 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 permitir a replicação e o 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.
- Standby 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 standby 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.
- Ativo/ativo de várias regiões (multissite) (RPO próximo a zero, RTO potencialmente zero): a workload é implantada em várias Regiões da AWS e processa ativamente o tráfego delas. Essa estratégia exige que você sincronize os dados entre regiões. Deve-se evitar ou lidar com 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 standby 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 standby 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 standby 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.

Quando o custo é uma preocupação e você deseja alcançar objetivos de RPO e RTO semelhantes, conforme definido na estratégia de standby passivo, é possível considerar soluções nativas da nuvem, como AWS Elastic Disaster Recovery, que adota a abordagem de luz piloto e oferece metas de RPO e RTO aprimoradas.

Etapas da implementação

1. Determine uma estratégia de DR que satisfaça os requisitos de recuperação para essa workload.

Escolher uma estratégia de DR é uma troca entre reduzir o tempo de inatividade e a perda de dados (RTO e RPO) e 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 o RTO máximo permitido e o orçamento limite da estratégia de restauração de serviço. Considerando os objetivos empresariais, as estratégias de DR luz-piloto e standby passivo satisfarão tanto o RTO quanto os critérios de custo.

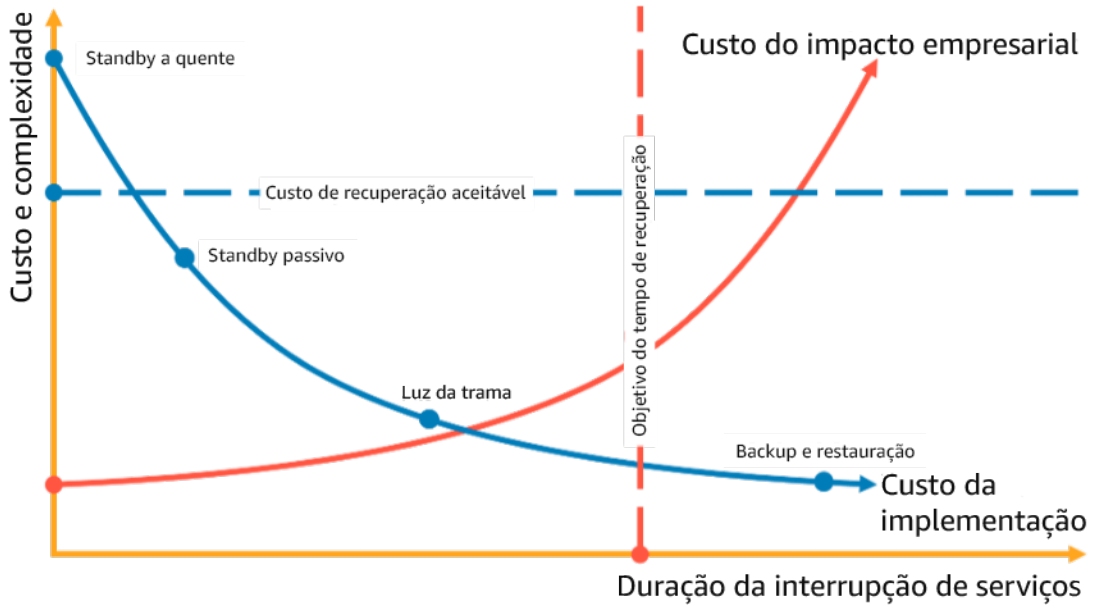


Figura 18: Escolha de uma estratégia de DR com base no RTO e no custo

Para saber mais, consulte Business Continuity Plan (BCP) (Plano de continuidade de negócios (BCP).

2. Revise os padrões de como a estratégia de DR selecionada pode ser implementada.

Essa 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 a seguir, é possível aplicar a estratégia para 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 dados e copiá-los para outro local (como outra Região da AWS).

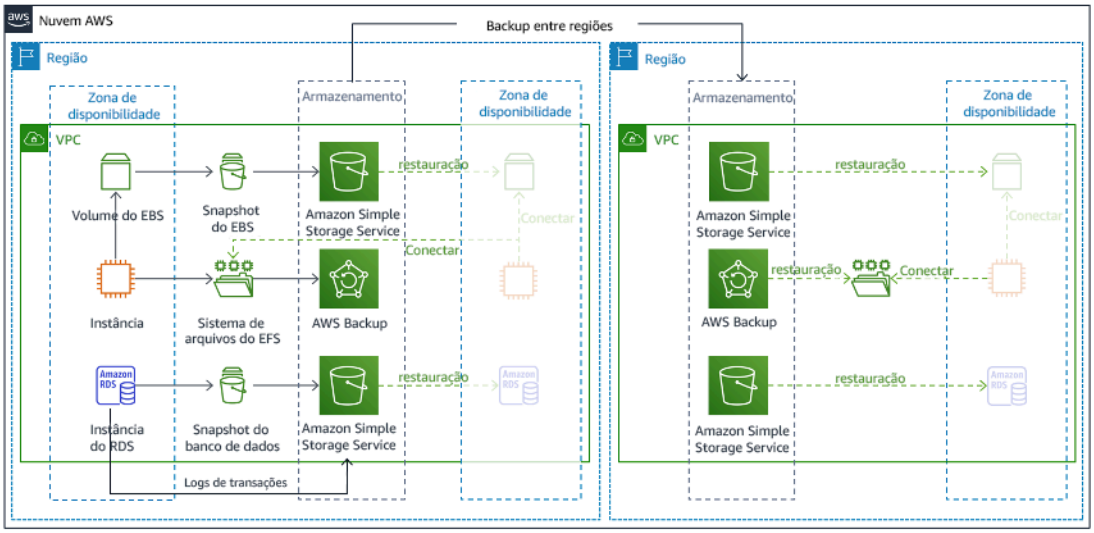


Figura 19: Arquitetura de backup e restauração

Para obter mais detalhes sobre essa estratégia, consulte Disaster Recovery (DR) Architecture on AWS, Part II: Backup and Restore with Rapid Recovery (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 principais recursos usados para a infraestrutura da workload são implantados na região de recuperação. No entanto, recursos adicionais e as dependências ainda são necessários para tornar a pilha funcional. Por exemplo, na Figura 20, nenhuma instância de computação é implantada.

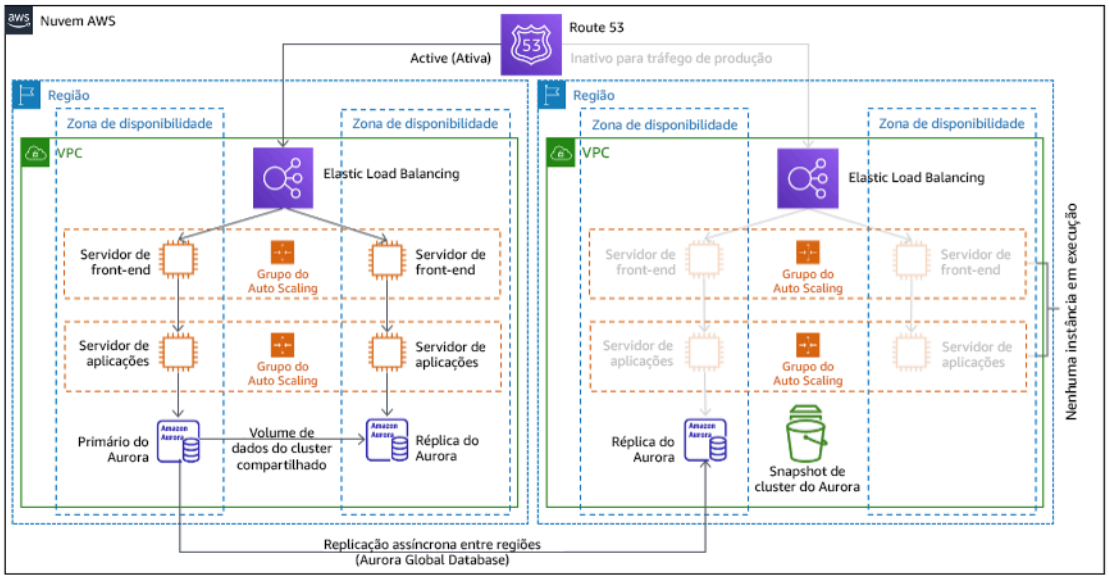


Figura 20: Arquitetura de luz-piloto

Para obter mais detalhes sobre essa estratégia, consulte [Disaster Recovery (DR) Architecture on AWS, Part III: Pilot Light and Warm Standby] (Arquitetura de recuperação de desastres (DR) na AWS, parte III: luz-piloto e standby passivo).

Modo de espera passivo

A abordagem de 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. Se a região de recuperação estiver implantada com sua capacidade total, isso é conhecido como standby a quente.

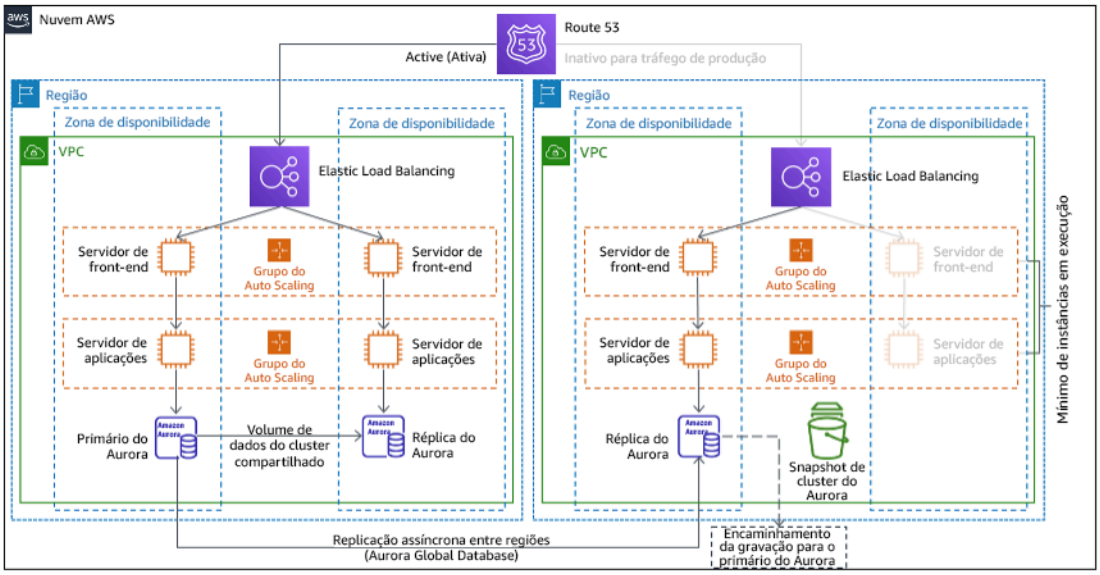


Figura 21: Arquitetura de standby passivo

O uso de standby passivo ou luz-piloto requer que a escala dos recursos seja aumentada verticalmente na região de recuperação. Para verificar se a capacidade está disponível quando necessário, considere o uso de reservas de capacidade para instâncias do EC2. Se você usar o AWS Lambda, a simultaneidade provisionada poderá fornecer ambientes de execução para que eles sejam preparados para responder imediatamente às invocações da função.

Para obter mais detalhes sobre essa estratégia, consulte Disaster Recovery (DR) Architecture on AWS, Part III: Pilot Light and Warm Standby (Arquitetura de recuperação de desastres (DR) na AWS, parte III: luz-piloto e standby passivo).

Multissite ativo/ativo

É possível executar a 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 essa estratégia por outros motivos, além da DR. Ela pode ser usada para aumentar a disponibilidade ou ao implantar uma workload para um público global (a fim de aproximar o endpoint dos usuários e/ou implantar pilhas localizadas para o público nessa região). Como uma estratégia de DR, se a workload não for compatível com uma das Regiões da AWS onde está implantada, essa 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 empresariais exigirem.

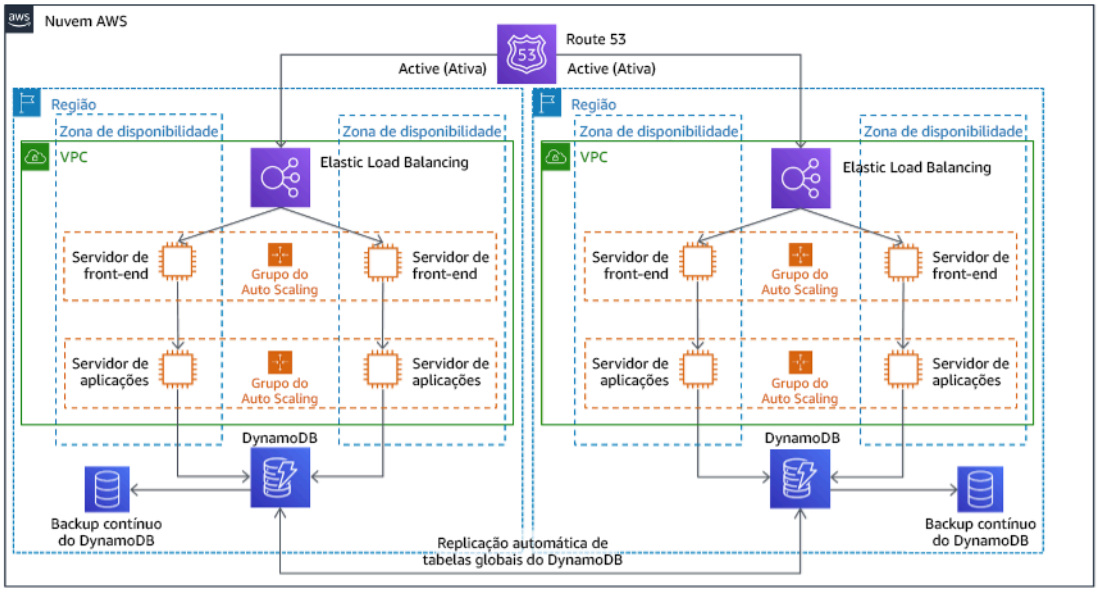


Figura 22: Arquitetura de multissite ativo-ativo

Para obter mais detalhes sobre essa estratégia, consulte Disaster Recovery (DR) Architecture on AWS, Part IV: Multi-site Active/Active (Arquitetura de recuperação de desastres (DR) na AWS, parte IV: multissite ativo/ativo).

AWS Elastic Disaster Recovery

Se você estiver considerando a estratégia de luz-piloto ou de standby passivo para a recuperação de desastres, o AWS Elastic Disaster Recovery poderia fornecer uma abordagem alternativa com benefícios melhorados. O Elastic Disaster Recovery pode oferecer uma meta de RPO e RTO semelhante à do standby passivo, mas mantém a abordagem de baixo custo da luz-piloto. O Elastic Disaster Recovery replica os dados da região primária para a região de recuperação, usando a proteção contínua de dados para alcançar um RPO medido em segundos e um RTO que pode ser medido em minutos. Somente os recursos necessários para replicar os dados são implantados na região de recuperação, o que mantém os custos baixos, semelhante à estratégia de luz-piloto. Ao usar o Elastic Disaster Recovery, o serviço coordena e orquestra a recuperação de recursos de computação quando iniciado como parte de um failover ou de uma simulação.

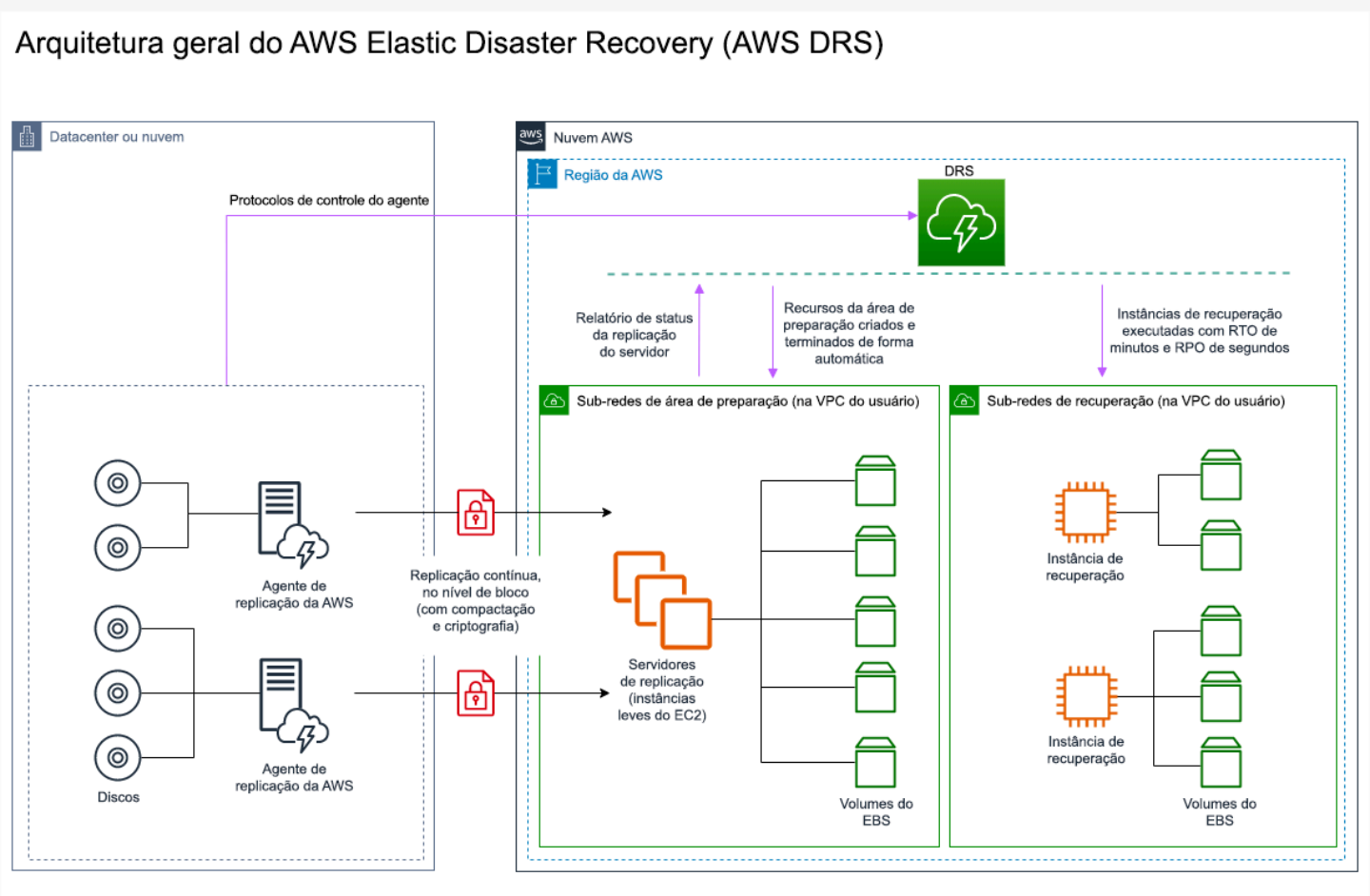


Figura 23: Arquitetura do AWS Elastic Disaster Recovery

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, a 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. Essa arquitetura usa uma abordagem multissite ativo/ativo, já que as instâncias do Amazon EC2 e o Elastic Load Balancer tem recursos implantados em várias AZs, processando solicitações ativamente. A arquitetura também demonstra o standby a

quente, no qual, caso a instância primária do Amazon RDS falhe (ou a própria AZ falhe), a instância de standby será promovida a primária.

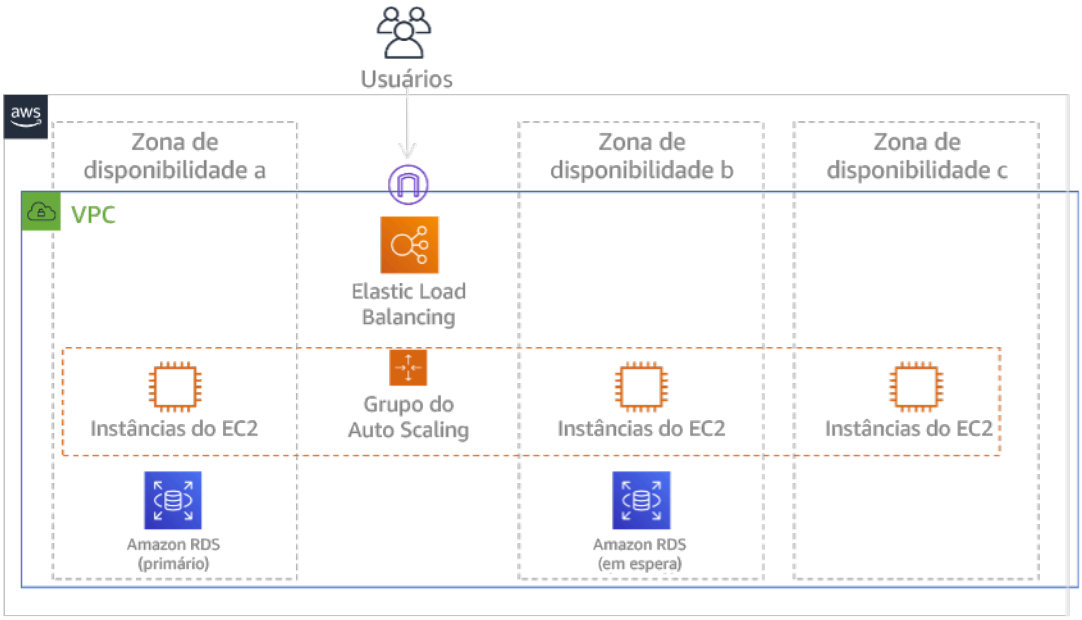


Figura 24: Arquitetura Multi-AZ

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 abordagem alternativa menos comum para uma única região, a DR Multi-AZ está ilustrada nesta publicação de blog, Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 1: Single-Region stack (Criar 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.

> **Note**
>
> 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 empresariais. 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 recursos de infraestrutura e 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 de luz-piloto e standby passivo, a infraestrutura implantada exigirá ações adicionais para ficar pronta para produção. Usando parâmetros e lógica condicional do CloudFormation, é possível controlar se uma pilha implantada está ativa ou em espera com um único modelo. Ao usar o Elastic Disaster Recovery, o serviço vai replicar e orquestrar a restauração de configurações da aplicação e os recursos de computação.

Todas as estratégias de DR exigem que seja feito backup das fontes de dados na Região da AWS e, então, esses backups são copiados para a região de recuperação. O AWS Backup fornece uma visão centralizada em que é possível configurar, programar e monitorar backups para esses recursos. Para luz-piloto, standby 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 regiões, consulte essa série de blog em Creating a Multi-Region Application with AWS Services (Criar uma aplicação de várias regiões 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 de luz-piloto e standby 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 da Amazon Virtual Private Cloud (Amazon VPC), de sub-redes e dos grupos de segurança necessários. É possível automatizar grande parte do processo de restauração. Para saber como, consulte esta 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 de forma manual ou automática. 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, geralmente é usado o failover iniciado manualmente. 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 Regiões 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 um failover iniciado manualmente, é possível usar o Amazon Route 53 Application Recovery Controller, o que fornece uma API de plano de dados altamente disponível para rotear novamente 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 essa e outras opções, consulte esta seção do whitepaper de recuperação de desastres.

6. Projete um plano de como será feito failback da workload.

Failback é quando você retorna a operação de workload para a região primária, após o término 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 você usar 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 você usar o banco de dados global do Amazon Aurora e o failover planejado e gerenciado, a topologia da replicação existente do banco de dados global do Aurora será mantida. Portanto, a antiga instância de leitura/gravação na região primária se tornará uma réplica e receberá atualizações da região de recuperação.

Em casos nos quais isso não seja 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 outras réplicas. Por exemplo, para obter instruções de como fazer isso com o banco de dados global do Amazon Aurora presumindo um failover não planejado, consulte este laboratório: Fail Back a Global Database (Executar failback em um banco de dados global).

Após um failover, se você puder continuar a execução 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 alternância programada, trocando as 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 manual disponível para todos os membros da equipe e que seja revisado periodicamente.

Ao usar o Elastic Disaster Recovery, o serviço auxiliará na orquestração e automatização do processo de failback. Para obter mais detalhes, consulte Performing a failback (Como executar failback).

Nível de esforço do 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: conceitos básicos – 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: padrões de arquitetura para aplicações ativas/ativas de várias regiões (ARC209-R2)
- Conceitos básicos do AWS Elastic Disaster Recovery | Amazon Web Services

Exemplos relacionados:

- Well-Architected Lab - Disaster Recovery - Series of workshops illustrating DR strategies (Laboratório do Well-Architected – Recuperação de desastres: série de workshops ilustrando estratégias de DR)

REL13-BP03 Testar a implementação da recuperação de desastres para validá-la

Teste regularmente o failover no site de recuperação para verificar se a operação está correta e que o RTO e o RPO sejam cumpridos.

Antipadrões comuns:

- Nunca execute failovers na produção.

Benefícios do estabelecimento dessa prática recomendada: testar regularmente seu plano de recuperação de desastres garante que ele funcione quando necessário e que sua equipe saiba como executar a estratégia.

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

Orientações para a implementação

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 nesse 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.

Etapas da implementação

1. Projete suas cargas de trabalho para recuperação. Teste regularmente seus caminhos de recuperação. A computação orientada para a recuperação identifica as características em sistemas que aprimoram a recuperação: 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 capacidade de reinicialização. Pratique o caminho da recuperação para verificar se é possível 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.
2. Para workloads com base no Amazon EC2, use o AWS Elastic Disaster Recovery para implementar e iniciar instâncias de simulação para a estratégia de DR. O AWS Elastic Disaster Recovery permite executar simulações com eficiência, o que ajuda você a se preparar para um evento de failover. Também é possível iniciar frequentemente as instâncias usando o Elastic Disaster Recovery para fins de teste e simulação sem redirecionar o tráfego.

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 Elastic Disaster Recovery
- Recuperação de desastres de workloads na AWS: recuperação na nuvem (whitepaper da AWS)
- AWS Elastic Disaster Recovery Preparing for Failover (AWS Elastic Disaster Recovery: preparação para failover)
- O projeto de computação orientado por recuperação de Berkeley/Stanford
- What is AWS Fault Injection Simulator? (O que e o AWS Fault Injection Simulator?)

Vídeos relacionados:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (AWS re:Invent 2018: padrões de arquitetura para aplicações ativas/ativas de várias regiões)
- AWS re:Invent 2019: Backup-and-restore and disaster-recovery solutions with AWS (AWS re:Invent 2019: soluções de backup e restauração e de recuperação de desastres com a AWS)

Exemplos relacionados:

- Well-Architected Lab - Testing for Resiliency (Laboratório do Well-Architected: teste 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

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. Você pode encontrar orientações prescritivas sobre implementação no Whitepaper sobre pilar de eficiência de performance.

Áreas de práticas recomendadas

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

Seleção de arquitetura

Perguntas

- PERFORMANCE 1. Como você seleciona os recursos e a arquitetura de nuvem apropriados para sua workload?

PERFORMANCE 1. Como você seleciona os recursos e a arquitetura de nuvem apropriados para sua workload?

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

Práticas recomendadas

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

PERF01-BP01 Conheça e compreenda os serviços e recursos de nuvem disponíveis

Continue a descobrir e aprender sobre serviços e configurações disponíveis que ajudam a tomar decisões e melhorar a eficiência da performance de suas workloads com base na arquietura.

Antipadrões comuns:

- Você usa a nuvem como um datacenter colocalizado.
- Você não moderniza sua aplicação após a migração para a nuvem.
- Você só usa um tipo de armazenamento para tudo que precisa ser mantido.
- Você usa tipos de instância mais próximos aos padrões atuais, no entanto, maiores, quando necessário.
- Você implanta e gerencia tecnologias disponíveis como serviços gerenciados.

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

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

Orientação para implementação

A AWS lança constantemente novos serviços e recursos que podem melhorar a performance e reduzir o custo das workloads na nuvem. Atualizar-se com relação a esses novos serviços e atributos

é crucial para manter a eficácia da performance na nuvem. Modernizar a arquitetura da workload também ajuda a acelerar a produtividade, impulsionar a inovação e ter acesso a mais oportunidades de crescimento.

Etapas da implementação

- Faça um inventário do software e da arquitetura usados para serviços relacionados a suas workloads. Decida sobre qual categoria de produtos você quer saber mais.
- Explore as ofertas da AWS para identificar e aprender sobre os serviços e as opções de configuração relevantes que podem ajudar você a melhorar a performance e reduzir os custos e a complexidade operacional.
  - Quais são as novidades da AWS?
  - Blog da AWS
  - AWS Skill Builder
  - Eventos e webinars da AWS
  - Treinamento da AWS and Certifications
  - Canal da AWS no Youtube
  - Workshops da AWS
  - Comunidades da AWS
- Use ambientes sandbox (sem produção) para aprender e experimentar novos serviços sem incorrer em custos adicionais.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS
- Crie aplicações modernas na AWS

Vídeos relacionados:

- This is my Architecture

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS

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

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

Antipadrões comuns:

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

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

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

Orientação para implementação

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

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

Etapas da implementação

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

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS
- AWS Enterprise Support

Vídeos relacionados:

- This is my Architecture

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS

PERF01-BP03 Inclua o custo nas decisões de arquitetura

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

Antipadrões comuns:

- Você só usa uma família de instâncias.
- Você não avalia soluções licenciadas em relação a soluções de código aberto.
- Você não define políticas de ciclo de vida de armazenamento.
- Você não analisa os novos serviços e recursos da Nuvem AWS.
- Você só usa o armazenamento em bloco.

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

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

Orientação para implementação

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

Etapas da implementação

- Estabeleça objetivos de custo, como limites orçamentários para a workload na nuvem.

- Identifique os principais componentes (como instâncias e armazenamento) que impulsionam o custo da workload. Você pode usar o AWS Pricing Calculator e o AWS Cost Explorer para identificar os principais fatores de custo na workload.

- Use práticas recomendadas de otimização de custos do Well-Architected para otimizar esses componentes principais em termos de custo.
- Monitore e analise constantemente os custos para identificar oportunidades de otimizar as workloads e economizar.
  - Use o AWS Budgets para receber alertas quando os custos forem inaceitáveis.
  - Use AWS Compute Optimizer ou AWS Trusted Advisor para receber recomendações de otimização de custos.
  - Use Detecção de Anomalias de Custos da AWS para obter detecção automática de anomalias de custo e análise da causa raiz.

Recursos

Documentos relacionados:

- A Detailed Overview of the Cost Intelligence Dashboard
- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS

Vídeos relacionados:

- This is my Architecture
- Optimize performance and cost for your AWS compute

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS
- Rightsizing with Compute Optimizer and Memory utilization enabled

- AWS Compute Optimizer Demo code

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

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

Antipadrões comuns:

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

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

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

Orientação para implementação

Identifique áreas críticas na arquitetura em termos de desempenho e impacto para o cliente. 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.

Etapas da implementação

- Entenda SLAs e requisitos de suas workloads.
- Defina claramente os fatores de avaliação. Os fatores podem estar relacionados a custo, confiabilidade, segurança e desempenho de suas workloads.
- Selecione arquitetura e serviços que possam atender às suas necessidades.

- Realize experiências e provas de conceitos (POCs) para avaliar os fatores e o impacto de certas trocas para os clientes e para a eficiência da arquitetura. Normalmente, workloads de alta disponibilidade, com bom desempenho e seguras consomem mais recursos da nuvem e, ao mesmo tempo, proporcionam uma melhor experiência ao cliente.

Recursos

Documentos relacionados:

- Amazon Builders' Library
- Amazon QuickSight KPIs
- Amazon CloudWatch RUM
- Documentação do X-Ray
- Understand resiliency patterns and trade-offs to architect efficiently in the cloud

Vídeos relacionados:

- Build a monitoring plan
- Optimize applications through Amazon CloudWatch RUM
- 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)

PERF01-BP05 Use políticas e arquiteturas de referência

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

Antipadrões comuns:

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

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

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

Orientação para implementação

Ter políticas internas na seleção de recursos e arquitetura fornece padrões e diretrizes a serem seguidos ao fazer escolhas arquitetônicas. Essas diretrizes simplificam o processo de tomada de decisão ao escolher o serviço de nuvem certo e podem ajudar a melhorar a eficiência da performance. Implante a workload usando políticas ou arquiteturas de referência. Integre os serviços à implantação na nuvem e, depois, use testes de desempenho para verificar se você pode continuar a atender aos seus requisitos de desempenho.

Etapas da implementação

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

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS

Vídeos relacionados:

- This is my Architecture

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS

PERF01-BP06 Use testes comparativos para orientar decisões de arquitetura

Compare o desempenho de uma workload existente para entender seu desempenho na nuvem e orientar decisões de arquitetura com base nesses dados.

Antipadrões comuns:

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

Benefícios de estabelecer esta prática recomendada: Os testes comparativos da implementação atual permitem medir a melhoria da performance.

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

Orientação para implementação

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

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

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

Como normalmente é mais rápido executar testes comparativos do que testes de carga, eles podem ser usados mais cedo no pipeline de implantação e fornecer um feedback mais rápido sobre desvios de performance. Ao avaliar uma alteração significativa em um componente ou serviço, um benchmark pode ser uma maneira rápida de verificar se é possível justificar a iniciativa para concretizar a alteração. O uso de testes comparativos em conjunto com testes de carga é importante porque o teste de carga informa como é a performance da workload em produção.

Etapas da implementação

- Defina as métricas (como utilização da CPU, latência ou throughput) para avaliar o desempenho da workload.
- Identifique e configure uma ferramenta de testes comparativos adequada à workload. Você pode usar serviços da AWS (como o Amazon CloudWatch) ou uma ferramenta de terceiros compatível com a workload.
- Execute testes comparativos e monitore as métricas durante o teste.
- Analise e documente os resultados do teste comparativo para identificar gargalos e problemas.
- Use os resultados do teste para tomar decisões de arquitetura e ajustar a workload. Isso pode incluir a mudança de serviços ou a adoção de novos recursos.
- Teste novamente a workload após o ajuste.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics

Vídeos relacionados:

- This is my Architecture
- Optimize applications through Amazon CloudWatch RUM

- Demo of Amazon CloudWatch Synthetics (Demonstração do Amazon CloudWatch Synthetics)

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS
- Distributed Load Tests
- 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)

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

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

Antipadrões comuns:

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

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

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

Orientação para implementação

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

Os atrasos de workloads críticas devem consistir não apenas em histórias de usuários que venham a oferecer funcionalidades relevantes para empresas e usuários, mas também em histórias técnicas

que formem uma base de arquitetura para as workloads. Essa base é formada por novos avanços em tecnologia e novos serviços e os adota com base em dados e justificativas adequadas. Isso verifica se a arquitetura permanece preparada para o futuro e não fica estagnada.

Etapas da implementação

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

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

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

Recursos

Documentos relacionados:

- Biblioteca de Soluções da AWS
- Central de Conhecimento da AWS

Vídeos relacionados:

- This is my Architecture

Exemplos relacionados:

- Amostras da AWS
- Exemplos de SDKs da AWS

Computação e hardware

PERFORMANCE 2. Como você seleciona e usa recursos computacionais em sua workload?

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

Práticas recomendadas

- PERF02-BP01 Selecione as melhores opções de computação para as workloads
- PERF02-BP02 Entenda a configuração e os recursos de computação disponíveis
- PERF02-BP03 Colete métricas relacionadas à computação
- PERF02-BP04 Configure e dimensione corretamente os recursos de computação
- PERF02-BP05 Dimensione recursos de computação dinamicamente
- PERF02-BP06 Use optimized hardware-based compute accelerators

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

Selecionar a opção de computação mais adequada para suas workloads permite que você melhore o desempenho, reduza os custos desnecessários de infraestrutura e reduza os esforços operacionais necessários para mantê-las.

Antipadrões comuns:

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

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

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

Orientação para implementação

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

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

Etapas da implementação

1. Entenda os requisitos de computação das workloads. Os principais requisitos a serem considerados incluem necessidades de processamento, padrões de tráfego, padrões de acesso a dados, necessidades de ajuste de escala e requisitos de latência.
2. Saiba mais sobre as diferentes opções de computação disponíveis para a workload na AWS (conforme descrito em PERF01-BP01 Conheça e compreenda os serviços e recursos de nuvem disponíveis. Veja algumas das principais opções de computação da AWS, as características e casos de uso comuns:

Serviço da AWS	Características principais	Casos de uso comum
Amazon Elastic Compute Cloud (Amazon EC2)	Tem opção dedicada para hardware, requisitos de licença, grande seleção de diferentes famílias de instâncias, tipos de processadores e acelerado res de computação.	Migrações do tipo mover sem alterações (lift-and-shift), aplicações monolíticas, ambientes híbridos, aplicaçõe s empresariais
Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	Fácil implantação, ambientes consistentes, escaláveis	Microsserviços, ambientes híbridos
AWS Lambda	Computação com tecnologi a sem servidor Serviço que executa código em resposta a eventos e gerencia automaticamente os recursos computacionais subjacentes.	Microsserviços, aplicações orientadas a eventos
AWS Batch	Provisiona e escala de forma eficiente e dinâmica. Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service	HPC, treine modelos de ML.

Serviço da AWS	Características principais	Casos de uso comum
	(Amazon EKS)e AWS Fargate Recursos de computação, com a opção de usar instâncias sob demanda ou spot com base nos requisitos de trabalho.	
Amazon Lightsail	Aplicação Linux e Windows pré-configurada para executar pequenas workloads .	Aplicações web simples, site personalizado.

3. Avalie o custo (como cobrança por hora ou transferência de dados) e as despesas gerais de gerenciamento (como aplicação de patches e ajuste de escala) associados a cada opção de computação.
4. Faça experimentos e análises comparativas em um ambiente que não seja de produção para identificar qual opção de computação pode melhor atender às necessidades da workload.
5. Depois de experimentar e identificar sua nova solução de computação, planeje a migração e valide as métricas de desempenho.
6. Use ferramentas de monitoramento da AWS, como Amazon CloudWatch e serviços de otimização, como AWS Compute Optimizer para otimizar continuamente os recursos de computação com base em padrões de uso do mundo real.

Recursos

Documentos relacionados:

- Cloud Compute with AWS
- Amazon EC2 Instance Types
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Amazon ECS Containers: Amazon ECS Container Instances
- Funções: configuração de função do Lambda
- Prescriptive Guidance for Containers
- Prescriptive Guidance for Serverless

Vídeos relacionados:

- How to choose compute option for startups (Como escolher uma opção de computação para startups)
- Optimize performance and cost for your AWS compute
- Amazon EC2 foundations
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Deploy ML models for inference at high performance and low cost
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2

Exemplos relacionados:

- Migrating the Web application to containers
- Run a Serverless Hello World

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

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

Antipadrões comuns:

- Você não avalia as opções de computação ou as famílias de instâncias disponíveis em relação às características da workload.
- Você provisiona recursos de computação em excesso para atender aos requisitos de pico de demanda.

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

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

Orientação para implementação

Cada solução de computação tem configurações e recursos exclusivos disponíveis para suportar diferentes características e requisitos das workloads. Saiba como essas opções complementam sua

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

Etapas da implementação

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

Opção de configuração	Exemplos
Tipo de instância	• As instâncias otimizadas para computação são ideais para workloads que exigem uma proporção maior de vCPU/memória. • As instâncias otimizadas para memória entregam grandes quantidades de memória para oferecer compatibilidade com as workloads com uso intenso de memória. • As instâncias otimizadas para armazenamento são projetadas para workloads que exigem alta leitura sequencial e acesso de gravação (IOPS) no armazenamento local.
Modelo de definição de preço	• 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 desempenho da linha de base. • Savings Plans oferecem economias significativas em relação às instâncias sob

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

Recursos

Documentos relacionados:

- Cloud Compute with AWS
- Amazon EC2 Instance Types
- Processor State Control for Your Amazon EC2 Instance
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Amazon ECS Containers: Amazon ECS Container Instances
- Funções: configuração de função do Lambda

Vídeos relacionados:

- Amazon EC2 foundations
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Optimize performance and cost for your AWS compute

Exemplos relacionados:

- Rightsizing with Compute Optimizer and Memory utilization enabled
- AWS Compute Optimizer Demo code

PERF02-BP03 Colete métricas relacionadas à computação

Registre e acompanhe métricas relacionadas à computação para entender melhor o desempenho de seus recursos e melhorar seu desempenho e utilização.

Antipadrões comuns:

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

Benefícios de estabelecer esta prática recomendada: A coleta de métricas relacionadas à performance ajudará você a alinhar a performance da aplicação aos requisitos empresariais para

garantir que você atenda às necessidades da workload. Isso também pode ajudar a melhorar constantemente o desempenho e a utilização dos recursos na workload.

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

Orientação para implementação

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

Etapas da implementação

1. Identifique quais métricas relacionadas ao desempenho são relevantes para a workload. Você deve coletar métricas sobre a utilização de recursos e a forma como a workload na nuvem está operando (como tempo de resposta e throughput).
   a. Métricas padrão do Amazon EC2
   b. Métricas padrão do Amazon ECS
   c. Métricas padrão do Amazon EKS
   d. Métricas padrão do Lambda
   e. Métricas de memória e disco do Amazon EC2
2. Escolha e configure a solução certa de registro e monitoramento para a workload.
   a. Observabilidade nativa da AWS
   b. AWS Distro para OpenTelemetry
   c. Amazon Managed Service for Prometheus
3. Defina o filtro e a agregação necessários para as métricas com base nos requisitos da workload.
   a. Quantify custom application metrics with Amazon CloudWatch Logs and metric filters
   b. Collect custom metrics with Amazon CloudWatch strategic tagging

4. Configure políticas de retenção de dados para que as métricas correspondam às metas operacionais e de segurança.
    a. Retenção de dados padrão para métricas do CloudWatch
    b. Retenção de dados padrão para o CloudWatch Logs
5. Se necessário, crie alarmes e notificações para as métricas a fim de ajudar a reagir proativamente a problemas relacionados à performance.
    a. Create alarms for custom metrics using Amazon CloudWatch anomaly detection
    b. Create metrics and alarms for specific web pages with Amazon CloudWatch RUM
6. Use a automação para implantar os agentes de agregação de métricas e logs.
    a. AWS Systems Manager Automation
    b. OpenTelemetry Collector

Recursos

Documentos relacionados:

- Documentação do Amazon CloudWatch
- Collect metrics and logs from Amazon EC2 instances and on-premises servers with the CloudWatch Agent
- Acessar o Amazon CloudWatch Logs para o AWS Lambda
- Using CloudWatch Logs with container instances
- Publicar métricas personalizadas
- AWS Answers: Centralized Logging
- AWS Services That Publish CloudWatch Metrics
- Monitoring Amazon EKS on AWS Fargate

Vídeos relacionados:

- Application Performance Management on AWS

Exemplos relacionados:

- Level 100: Monitoring with CloudWatch Dashboards

- Level 100: Monitoring Windows EC2 instance with CloudWatch Dashboards
- Level 100: Monitoring an Amazon Linux EC2 instance with CloudWatch Dashboards

PERF02-BP04 Configure e dimensione corretamente os recursos de computação

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

Antipadrões comuns:

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

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

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

Orientação para implementação

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

Etapas da implementação

- Escolha um tipo de instância que melhor atenda às suas necessidades:
  - Como faço para escolher o tipo de instância do Amazon EC2 apropriado para minha workload?
  - Seleção de tipo de instância baseada em atributos para frota do Amazon EC2
  - Create an Auto Scaling group using attribute-based instance type selection
  - Optimizing your Kubernetes compute costs with Karpenter consolidation (Otimizar seus custos de computação do Kubernetes com a consolidação do Karpenter)
- 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 que melhor correspondam ao perfil e às metas de desempenho da workload.
- Monitore o uso de recursos usando ferramentas de monitoramento da AWS, como o Amazon CloudWatch.
- Selecione a configuração correta para os recursos computacionais.
  - Para workloads efêmeras, avalie métricas do Amazon CloudWatch para instâncias , como `CPUUtilization` para identificar se a instância está subutilizada ou superutilizada.
  - Para workloads estáveis, verifique as ferramentas de dimensionamento correto da AWS, como AWS Compute Optimizer e AWS Trusted Advisor em intervalos regulares, para identificar oportunidades de otimizar e dimensionar corretamente o recurso de computação.
    - Laboratório do Well-Architected: Recomendações de dimensionamento correto
    - Laboratório do Well-Architected: Dimensionamento correto com o Compute Optimizer
- Teste as alterações na configuração em um ambiente que não seja de produção antes de implementá-las em um ambiente ativo.
- Reavalie constantemente novas ofertas de computação e as compare com as necessidades da workload.

Recursos

Documentos relacionados:

- Cloud Compute with AWS
- Amazon EC2 Instance Types
- Amazon ECS Containers: Amazon ECS Container Instances
- Amazon EKS Containers: Amazon EKS Worker Nodes

- Funções: configuração de função do Lambda
- Processor State Control for Your Amazon EC2 Instance

Vídeos relacionados:

- Amazon EC2 foundations
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2
- Deploy ML models for inference at high performance and low cost
- Optimize performance and cost for your AWS compute
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Como simplificar o processamento de dados para aprimorar a inovação com ferramentas de tecnologia sem servidor

Exemplos relacionados:

- Rightsizing with Compute Optimizer and Memory utilization enabled
- AWS Compute Optimizer Demo code

PERF02-BP05 Dimensione recursos de computação dinamicamente

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

Antipadrões comuns:

- Você reage a alarmes aumentando a capacidade manualmente.
- Você usa as mesmas diretrizes de dimensionamento (geralmente infraestrutura estática) do ambiente on-premises.
- Você deixa a capacidade aumentada após um evento de escalabilidade, em vez de reduzir novamente.

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

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

Orientação para implementação

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

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

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

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

Etapas da implementação

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

Mecanismo de ajuste de escala automático	Onde usar
Amazon EC2 Auto Scaling	Para garantir que você tenha o número correto de instâncias do Amazon EC2 disponíveis para lidar com a carga do usuário para a aplicação.

Mecanismo de ajuste de escala automático	Onde usar
Application Auto Scaling	Para escalar automaticamente os recursos para serviços individuais da AWS além do Amazon EC2, como funções do AWS Lambda ou os serviços Amazon Elastic Container Service (Amazon ECS) .
Kubernetes Cluster Autoscaler/Karpenter	Para escalar automaticamente os clusters do Kubernetes.

- O ajuste de escala geralmente é discutido em relação a serviços de computação, como instâncias do Amazon EC2 ou funções do AWS Lambda. Não se esqueça de pensar também na configuração de serviços não computacionais, como AWS Glue para atender à demanda.
- Verifique se as métricas de ajuste de escala correspondem às características da workload que está sendo implantada. Se você estiver implantando uma aplicação de transcodificação de vídeo, espera-se que a utilização da CPU seja de 100%, e essa não deve ser sua métrica principal. Em vez disso, use a profundidade da fila de trabalhos de transcodificação. Você pode usar uma métrica personalizada 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.
- Use 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 da workload podem lidar com os dois eventos de ajuste de escala (aumento e redução). Por exemplo, você pode usar o histórico de atividades para 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. Com a escalabilidade preditiva, é possível eliminar a necessidade de superprovisionar a capacidade. Para obter mais detalhes, consulte Ajuste de escala com o Amazon EC2 Auto Scaling.

Recursos

Documentos relacionados:

- Cloud Compute with AWS
- Amazon EC2 Instance Types
- Amazon ECS Containers: Amazon ECS Container Instances
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Funções: configuração de função do Lambda
- Processor State Control for Your Amazon EC2 Instance
- Deep Dive on Amazon ECS Cluster Auto Scaling
- Introducing Karpenter – An Open-Source High-Performance Kubernetes Cluster Autoscaler (Apresentação do Karpenter: um dimensionador automático de clusters do Kubernetes de código aberto e alto desempenho)

Vídeos relacionados:

- Amazon EC2 foundations
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2
- Optimize performance and cost for your AWS compute
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Criar um ambiente de computação eficiente em termos de custo, energia e recursos

Exemplos relacionados:

- Amazon EC2 Auto Scaling Group Examples
- Implement Autoscaling with Karpenter

PERF02-BP06 Use optimized hardware-based compute accelerators

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

Antipadrões comuns:

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

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

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

Orientação para implementação

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

Etapas da implementação

- Identifique quais instâncias com computação acelerada podem atender aos seus requisitos.
- Para workloads de machine learning, utilize hardware específico para sua workload, como AWS Trainium, AWS Inferentiae o Amazon EC2 DL1. Instâncias do AWS Inferentia, como instâncias Inf2, oferecem até 50% melhor performance/watt em relação a instâncias comparáveis do Amazon EC2.
- Colete métricas de uso para as instâncias com computação acelerada. Por exemplo, você pode usar o agente do CloudWatch para coletar métricas como `utilization_gpu` e `utilization_memory` para suas GPUs, conforme mostrado em Colete métricas da GPU NVIDIA com o Amazon CloudWatch.
- Otimize o código, a operação de rede e as configurações dos aceleradores de hardware para garantir que o hardware subjacente seja totalmente utilizado.
  - Otimizar as configurações da GPU
  - Monitoramento e otimização de GPU no Deep Learning AMI

  - Otimização de E/S para ajuste de desempenho de GPU de treinamento de aprendizado profundo no Amazon SageMaker
- Use as mais recentes bibliotecas de alto desempenho e drivers de GPU.
- Use automação para liberar instâncias de GPU quando não estiverem em uso.

Recursos

Documentos relacionados:

- Instâncias de GPU
- Instâncias com AWS Trainium
- Instâncias com o AWS Inferentia
- Let's Architect! Arquitetura com chips e aceleradores personalizados

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

Vídeos relacionados:

- How to select Amazon EC2 GPU instances for deep learning (Como selecionar instâncias de GPU do Amazon EC2 para aprendizado profundo)
- Deploying Cost-Effective Deep Learning Inference (Implantação de inferência de aprendizado profundo econômico)

Gerenciamento de dados

PERFORMANCE 3. Como você armazena, gerencia e acessa dados em sua workload?

A solução de gerenciamento de dados ideal para um sistema específico varia conforme o tipo de dados (bloco, arquivo ou objeto), os padrões de acesso (aleatório ou sequencial), o throughput necessário, a frequência de acesso (online, offline, arquivamento), a frequência de atualização

(WORM, dinâmica) e as restrições de disponibilidade e durabilidade. As workloads do Well-Architected usam datastores específicos que permitem que recursos diferentes melhorem o desempenho.

Práticas recomendadas

- PERF03-BP01 Use um armazenamento de dados específico que melhor atenda aos seus requisitos de acesso e armazenamento de dados
- PERF03-BP02 Avalie as opções de configuração disponíveis para o datastore
- PERF03-BP03 Colete e registre métricas de desempenho do datastore
- PERF03-BP04 Implemente estratégias para melhorar o desempenho da consulta no datastore
- PERF03-BP05 Implementar padrões de acesso a dados que utilizem cache

PERF03-BP01 Use um armazenamento de dados específico que melhor atenda aos seus requisitos de acesso e armazenamento de dados

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

Antipadrões comuns:

- Fixar-se em um único datastore porque há experiência e conhecimento internos de um tipo específico de solução de banco de dados.
- Você pressupõe que todas as workloads têm requisitos de acesso e armazenamento de dados semelhantes.
- Você não implementou um catálogo de dados para criar um inventário de seus ativos de dados.

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

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

Orientação para implementação

Ao selecionar e implementar o armazenamento de dados, certifique-se de que as características de consulta, ajuste de escala e armazenamento atendam aos requisitos de dados da workload.

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

Etapas da implementação

- Realize um inventário dos vários tipos de dados que existem na workload.
- Entenda e documente as características e os requisitos dos dados, incluindo:
  - Tipo de dados (não estruturados, semiestruturados, relacionais)
  - Volume e crescimento de dados
  - Durabilidade dos dados: persistentes, efêmeros, transitórios
  - Requisitos de ACID (atomicidade, consistência, isolamento, durabilidade)
  - Padrões de acesso a dados (com muita leitura ou gravação)
  - Latência
  - Taxa de transferência
  - IOPS (operações de entrada/saída por segundo)
  - Período de retenção de dados
- Conheça os diferentes datastores disponíveis para a workload na AWS que podem atender às características dos dados (conforme descrito em PERF01-BP01 Conheça e compreenda os serviços e recursos de nuvem disponíveis). Alguns exemplos de tecnologias de armazenamento da AWS e suas principais características incluem:

Tipo	Serviços da AWS	Características principais
Armazenamento de objetos	Amazon S3	Escalabilidade ilimitada, alta disponibilidade e várias opções de acessibilidade. A transferência e o acesso a objetos dentro e fora do Amazon S3 podem usar um

Tipo	Serviços da AWS	Características principais
		serviço, como Aceleração de Transferências ou Pontos de Acesso, para oferecer compatibilidade com o local, necessidades de segurança e padrões de acesso.
Armazenamento de arquivamento	Amazon S3 Glacier	Desenvolvido para arquivamento de dados.
Armazenamento de streaming	Amazon Kinesis Amazon Managed Streaming for Apache Kafka (Amazon MSK)	Ingestão e armazenamento eficientes de dados de streaming.
Sistema de arquivos compartilhado	Amazon Elastic File System (Amazon EFS)	Sistema de arquivos montável que pode ser acessado por vários tipos de soluções de computação.

Tipo	Serviços da AWS	Características principais
Sistema de arquivos compartilhado	Amazon FSx	Baseia-se nas soluções de computação mais recentes da AWS para oferecer compatibilidade com quatro sistemas de arquivos usados com frequência: NetApp ONTAP, OpenZFS, Windows File Server e Lustre. Amazon FSx latência, throughput e IOPS 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.
O Armazenamento em bloco	Amazon Elastic Block Store (Amazon EBS)	Serviço de armazenamento de bloco escalável e de alta performance projetado para Amazon Elastic Compute Cloud (Amazon EC2). O Amazon EBS inclui armazenamento com base em SSD para workloads transacionais e de uso intenso de IOPS e armazenamento com base em HDD para workloads de uso intenso de throughput.

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

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

Tipo	Serviços da AWS	Características principais
Banco de dados de grafos	Amazon Neptune	Utilizado para aplicações que precisam navegar e consultar milhões de relacionamentos entre conjuntos de dados de grafos altamente 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.
Banco de dados de séries temporais	Amazon Timestream	Utilizado para coletar, sintetizar e gerar com eficiência insights de dados que mudam ao longo do tempo. Aplicativos de IoT, DevOps e telemetria industrial podem utilizar bancos de dados de séries temporais.
Coluna ampla	Amazon Keyspaces (para Apache Cassandra)	Usa tabelas, linhas e colunas, mas ao contrário de um banco de dados relacional, os nomes e o formato das colunas podem variar de linha para linha na mesma tabela. Normalmente, você vê um repositório de coluna ampla em aplicativos industriais de alta escala para manutenção de equipamentos, gerenciamento de frotas e otimização de rotas.

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

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

Pergunta	Fatos a serem considerados
Como os dados são estruturados?	• Se os dados estiverem estruturados, pense em um armazenamento de objetos, como o Amazon S3 , ou um banco de dados NoSQL, como o Amazon DocumentDB • Para dados de chave-valor, pense no DynamoDB, o Amazon ElastiCache for Redis ou Amazon MemoryDB for Redis
Qual nível de integridade referencial é necessário?	• Para restrições de chave externa, bancos de dados relacionais, como o Amazon RDS e o Aurora , podem oferecer esse nível de integridade.

Pergunta	Fatos a serem considerados
	• 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 em vez de unir documentos ou tabelas de diferentes locais.
A conformidade com ACID (atomicidade, consistência, isolamento, durabilidade) é necessária?	• Se as propriedades ACID associadas aos bancos de dados relacionais forem necessárias, pense em um banco de dados relacional, como o Amazon RDS e o Aurora. • Se for necessária uma consistência forte para banco de dados NoSQL, você pode usar leituras altamente consistentes com DynamoDB.
Como as necessidades de armazenamento serão alteradas ao longo do tempo? Como isso afeta a escalabilidade?	• Bancos de dados de tecnologia sem servidor como o DynamoDB e o Amazon Quantum Ledger Database (Amazon QLDB) serão escalados dinamicamente. • Os bancos de dados relacionais têm limites superiores em armazenamento provisionado e devem ser particionados horizontalmente usando mecanismos, como fragmentação, quando atingem esses limites.

Pergunta	Fatos a serem considerados
Qual é a proporção de consultas de leitura em relação a consultas de gravação? O armazenamento em cache melhoraria a performance?	• Workloads de uso intenso de leitura podem se beneficiar de uma camada de armazenamento em cache, como o ElastiCache ou DAX se o banco de dados for o DynamoDB. • As leituras também podem ser descarreg adas em réplicas de leitura com bancos de dados relacionais, como o Amazon RDS.
O armazenamento e a modificação (OLTP – Processamento de transações on-line) ou a recuperação e a geração de relatórios (OLAP – Processamento analítico on-line) têm uma prioridade mais alta?	• Para um processamento transacional de throughput alto de leitura no estado em que se encontra, considere um banco de dados NoSQL, como o DynamoDB. • Para padrões de leitura complexos e de throughput alto (como junção) com consistência use o Amazon RDS. • Para consultas de análise, pense em um banco de dados em colunas, como o Amazon Redshift , ou exporte os dados para o Amazon S3 e realize análises usando o Athena ou Amazon QuickSight.

Pergunta	Fatos a serem considerados
Qual nível de durabilidade os dados exigem?	• O Aurora replica automaticamente os dados entre três zonas de disponibilidade em uma região, o que significa que seus dados terão mais durabilidade com menos chance de serem perdidos. • O DynamoDB é automaticamente replicado entre várias zonas de disponibilidade, fornecendo alta disponibilidade e durabilid ade aos dados. • O Amazon S3 fornece 11 noves de durabilid ade. Muitos serviços de banco de dados, como o Amazon RDS e o DynamoDB, são compatíveis com a exportação de dados ao Amazon S3 para retenção de longo prazo e arquivamento.
Você quer se livrar de mecanismos de bancos de dados comerciais ou custos de licenças?	• Considere os mecanismos de código aberto, como o PostgreSQL e o MySQL no Amazon RDS ou no Aurora. • Utilize o AWS Database Migration Service e o AWS Schema Conversion Tool para realizar migrações de mecanismos de bancos de dados comerciais para código aberto
Qual a expectativa operacional para o banco de dados? A mudança para serviços gerenciados é uma preocupação principal?	• Utilizar o Amazon RDS em vez do Amazon EC2 e o DynamoDB ou o Amazon DocumentDB em vez de um host automátic o de um banco de dados NoSQL pode reduzir a sobrecarga operacional.

Pergunta	Fatos a serem considerados
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?	• 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.

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

Recursos

Documentos relacionados:

- Tipos de volume do Amazon EBS
- Armazenamento do Amazon EC2
- Amazon EFS: performance do Amazon EFS
- Performance do Amazon FSx for Lustre
- Performance do Amazon FSx for Windows File Server
- Documentação do Amazon S3 Glacier: S3 Glacier
- Amazon S3: considerações sobre performance e taxa de solicitação
- Armazenamento na nuvem com a AWS
- Características de E/S do Amazon EBS
- Bancos de dados em nuvem com a AWS
- Armazenamento em cache de banco de dados da AWS
- 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
- Choose between Amazon EC2 and Amazon RDS
- Melhores práticas para a implementação do Amazon ElastiCache

Vídeos relacionados:

- Deep dive on Amazon EBS
- Optimize your storage performance with Amazon S3
- Modernize apps with purpose-built databases
- Amazon Aurora storage demystified: How it all works
- Amazon DynamoDB deep dive: Advanced design patterns

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
- Optimize Data Pattern using Amazon Redshift Data Sharing
- Migrações de bancos de dados
- MS SQL Server - AWS Database Migration Service (AWS DMS) Replication Demo
- Database Modernization Hands On Workshop (Workshop prático de modernização de bancos de dados)
- Amostras da Amazon Neptune

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

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

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 não tem ciência das opções de configuração da solução de gerenciamento de dados escolhida.
- Você conta somente com o aumento do tamanho da instância sem examinar outras opções de configuração.
- Você não testa as características de ajuste de escala do datastore.

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

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

Orientação para implementação

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

Etapas da implementação

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

Opção de configuração	Exemplos
Offloading reads (like read replicas and caching)	• Em tabelas do DynamoDB, é possível descarregar leituras usando o DAX para armazenamento em cache. • Você pode criar um cluster do Amazon ElastiCache for Redis e configurar a aplicação para ler primeiro do cache e voltar para o banco de dados caso o item solicitado não esteja presente. • Todos os bancos de dados relacionais, como Amazon RDS e Aurora, e bancos de dados NoSQL provisionados, como Neptune e Amazon DocumentDB, permitem adicionar réplicas de leitura para descarregar as partes de leitura da workload. • Os bancos de dados de tecnologia sem servidor, como o DynamoDB, ajustarão a escala automaticamente. Verifique se você tem unidades de capacidade de leitura (RCU) suficientes provisionadas para processar a workload.

Opção de configuração	Exemplos
Scaling writes (like partition key sharding or introducing a queue)	• 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 no armazenamento subjacente. • Também é possível introduzir uma fila na frente do banco de dados, em vez de gravar diretamente nele. Esse padrão permite desacoplar a ingestão do banco de dados e controlar a taxa de fluxo, para que o banco de dados não fique sobrecarregado. • Usar solicitações de gravação em lote em vez de criar muitas transações de curta duração pode ajudar a melhorar o throughput em bancos de dados relacionais de alto volume de gravação. • Os bancos de dados de tecnologia sem servidor, como o DynamoDB, podem ajustar a escala do throughput de gravação automaticamente ou ajustar as unidades da capacidade de gravação (WCU) provisionadas, dependendo do modo da capacidade. • Você ainda pode ter problemas com partições ativas ao atingir os limites de throughput de 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.

Opção de configuração	Exemplos
Policies to manage the lifecycle of your datasets	• Você pode usar o Ciclo de Vida do Amazon S3 para gerenciar os objetos em todo o ciclo de vida. Se os padrões de acesso forem desconhecidos, variáveis ou imprevisíveis, você pode usar o Amazon S3 Intelligent-Tiering, que monitora padrões de acesso e move automaticamente objetos que não foram acessados para níveis de acesso de menor custo. Você pode aproveitar as métricas de Lente de Armazenamento do Amazon S3 para identificar oportunidades de otimização e lacunas no gerenciamento do ciclo de vida. • Gerenciamento do ciclo de vida do Amazon EFS gerencia automaticamente o armazenamento de arquivos para os sistemas de arquivos.
Gerenciamento de conexões e agrupamento	• O Amazon RDS Proxy pode ser usado com o Amazon RDS e o Aurora para gerenciar as conexões com o banco de dados. • 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 ajuste de escala automático para lidar com picos na carga.

- Realize experimentos e testes comparativos em um ambiente que não seja de produção para identificar qual opção de configuração pode atender aos requisitos da workload.
- Depois de experimentar, planeje a migração e valide as métricas de desempenho.
- Use ferramentas de monitoramento da AWS (como o Amazon CloudWatch) e otimização (como a Lente de Armazenamento do Amazon S3) para otimizar continuamente o armazenamento de dados usando o padrão de uso do mundo real.

Recursos

Documentos relacionados:

- Armazenamento na nuvem com a AWS
- Tipos de volume do Amazon EBS
- Armazenamento do Amazon EC2
- Amazon EFS: performance do Amazon EFS
- Performance do Amazon FSx for Lustre
- Performance do Amazon FSx for Windows File Server
- Documentação do Amazon S3 Glacier: S3 Glacier
- Amazon S3: considerações sobre performance e taxa de solicitação
- Armazenamento na nuvem com a AWS
- Armazenamento na nuvem com a AWS
- Características de E/S do Amazon EBS
- Bancos de dados em nuvem com a AWS
- Armazenamento em cache de banco de dados da AWS
- 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:

- Deep dive on Amazon EBS
- Optimize your storage performance with Amazon S3
- Modernize apps with purpose-built databases
- Amazon Aurora storage demystified: How it all works
- Amazon DynamoDB deep dive: Advanced design patterns

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
- Exemplos do Amazon DynamoDB
- AWS Database migration samples
- Database Modernization Workshop (Workshop de modernização de bancos de dados)
- Working with parameters on your Amazon RDS for Postgress DB

PERF03-BP03 Colete e registre métricas de desempenho do datastore

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

Antipadrões comuns:

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

Benefícios de estabelecer esta prática recomendada: O estabelecimento de uma linha de base de performance ajuda a compreender o comportamento normal e os requisitos das workloads. Padrões anormais podem ser identificados e depurados mais rapidamente, melhorando a performance e a confiabilidade do datastore.

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

Orientação para implementação

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

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

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

Etapas da implementação

1. Identifique as principais métricas de desempenho que o datastore deve monitorar.
    a. Métricas e dimensões do Amazon S3
    b. Métricas de monitoramento para em uma instância do Amazon RDS
    c. Monitorar a carga do banco de dados com o Performance Insights no Amazon RDS
    d. Visão geral do monitoramento aprimorado
    e. Métricas e dimensões do DynamoDB
    f. Monitoramento do DynamoDB Accelerator
    g. Monitoramento do Amazon MemoryDB for Redis com o Amazon CloudWatch
    h. Quais métricas devo monitorar?
    i. Monitoramento da performance do cluster do Amazon Redshift
    j. Métricas e dimensões do Timestream
    k. Métricas do Amazon CloudWatch para Amazon Aurora
    l. Registro em log e monitoramento no Amazon Keyspaces (for Apache Cassandra)
    m. Monitoramento dos recursos do Amazon Neptune
2. Use uma solução aprovada de registro em log e monitoramento para coletar essas métricas. 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.

3. Confira se o monitoramento do datastore pode se beneficiar de uma solução de machine learning que detecta anomalias de performance.
   a. O Amazon DevOps Guru para Amazon RDS fornece visibilidade dos problemas de performance e faz recomendações de ações corretivas.
4. Configure a retenção de dados em sua solução de monitoramento e registro para corresponder às suas metas operacionais e de segurança.
   a. Retenção de dados padrão para métricas do CloudWatch
   b. Retenção de dados padrão para o CloudWatch Logs

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
- 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
- Amazon Aurora storage demystified: How it all works
- Amazon DynamoDB deep dive: Advanced design patterns
- Best Practices for Monitoring Redis Workloads on Amazon ElastiCache

Exemplos relacionados:

- Level 100: Monitoring with CloudWatch Dashboards
- AWS Dataset Ingestion Metrics Collection Framework
- Amazon RDS Monitoring Workshop

PERF03-BP04 Implemente estratégias para melhorar o desempenho da consulta no datastore

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

Antipadrões comuns:

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

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

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

Orientação para implementação

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

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

Etapas da implementação

- Entenda e analise as consultas críticas de dados que são realizadas no datastore.
- Identifique as consultas com execução lenta no datastore e use planos de consulta para entender o estado atual delas.

  - Analyzing the query plan in Amazon Redshift
  - Using EXPLAIN and EXPLAIN ANALYZE in Athena
- Implemente estratégias para melhorar o desempenho da consulta. Algumas das principais estratégias incluem:
  - Usar um formato de arquivo colunar (como Parquet ou ORC).
  - Compactar os dados no datastore para reduzir o espaço de armazenamento e a operação de E/S.
  - Particionar os dados para dividi-los em partes menores e reduzir o tempo de verificação dos dados.
    - Partitioning data in Athena
    - Partições e distribuição de dados
  - Indexação de dados nas colunas comuns na consulta.
  - Escolha a operação de junção correta para consulta. Ao unir duas tabelas, especifique a tabela maior no lado esquerdo da junção e a tabela menor no lado direito.
  - Solução de cache distribuído para melhorar a latência e reduzir o número de operações de E/S do banco de dados.
  - Manutenção regular, como execução de estatísticas.
- Experimente e teste estratégias em um ambiente que não seja de produção.

Recursos

Documentos relacionados:

- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- 10 melhores dicas de desempenho do Amazon Athena
- Armazenamento em cache de banco de dados da AWS
- Melhores práticas para a implementação do Amazon ElastiCache
- Partitioning data in Athena

Vídeos relacionados:

- Optimize Data Pattern using Amazon Redshift Data Sharing

- Optimize Amazon Athena Queries with New Query Analysis Tools

Exemplos relacionados:

- Driver CSI do Amazon EFS

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

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

Antipadrões comuns:

- Você armazena em cache dados que mudam com frequência.
- Você depende dos dados em cache como se estivessem armazenados de forma durável e sempre disponíveis.
- Você não leva em conta a consistência dos seus dados em cache.
- Você não monitora a eficiência da sua implementação de cache.

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

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

Orientação para implementação

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

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

Armazenamento em cache do lado do cliente

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

Uma desvantagem dessa abordagem é que, quando vários clientes estão envolvidos, eles podem armazenar os mesmos dados em cache localmente. Isso resulta no uso de armazenamento duplicado e na inconsistência de dados entre esses clientes. Um cliente pode armazenar em cache os resultados de uma consulta e, um minuto depois, outro cliente pode executar a mesma consulta e obter um resultado diferente.

Armazenamento em cache remoto

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

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

Etapas da implementação

1. Identifique bancos de dados, APIs e serviços de rede que poderiam se beneficiar do armazenamento em cache. Serviços que têm workloads de leitura pesadas, uma alta taxa de leitura e gravação ou que são caros para escalar são candidatos ao armazenamento em cache.
   - Armazenamento em cache de banco de dados
   - Ativação do armazenamento em cache da API para melhorar a capacidade de resposta

2. Identifique o tipo apropriado de estratégia de armazenamento em cache que melhor se adapte ao seu padrão de acesso.
   - Estratégias de armazenamento em cache
   - Soluções de armazenamento em cache da AWS
3. Siga Práticas recomendadas de armazenamento em cache para seu armazenamento de dados.
4. Configure uma estratégia de invalidação de cache, como um time-to-live (TTL), para todos os dados que equilibre a atualização dos dados e reduza a pressão sobre o datastore de back-end.
5. Ative recursos como novas tentativas automáticas de conexão, recuo exponencial, tempos limite do lado do cliente e pool de conexões no cliente, se disponíveis, pois eles podem melhorar o desempenho e a confiabilidade.
   - Práticas recomendadas: clientes Redis e Amazon ElastiCache for Redis
6. Monitore a taxa de acertos de cache com uma meta de 80% ou mais. Valores mais baixos podem indicar tamanho insuficiente do cache ou um padrão de acesso que não se beneficia do armazenamento em cache.
   - Quais métricas devo monitorar?
   - Práticas recomendadas para monitorar workloads do Redis no Amazon ElastiCache
   - Monitoramento das práticas recomendadas com Amazon ElastiCache for Redis usando o Amazon CloudWatch
7. Implemente replicação de dados para descarregar as leituras em várias instâncias e melhorar o desempenho e a disponibilidade da leitura de dados.

Recursos

Documentos relacionados:

- Uso do Amazon ElastiCache Well-Architected Lens
- Monitoramento das práticas recomendadas com Amazon ElastiCache for Redis usando o Amazon CloudWatch
- Quais métricas devo monitorar?
- Whitepaper: Performance at Scale with Amazon ElastiCache (Desempenho em escala com Amazon ElastiCache)
- Desafios e estratégias de armazenamento em cache

Vídeos relacionados:

- Amazon ElastiCache Learning Path (Roteiro de aprendizado do Amazon ElastiCache)
- Design for success with Amazon ElastiCache best practices (Projete para o sucesso com as práticas recomendadas do Amazon ElastiCache)

Exemplos relacionados:

- Como aumentar o desempenho do banco de dados MySQL com Amazon ElastiCache for Redis

Rede e entrega de conteúdo

PERFORMANCE 4. Como você seleciona e configura os recursos de rede em sua workload?

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

PERF04-BP01 Compreender como as redes afetam a performance

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

Antipadrões comuns:

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

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

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

Orientação para implementação

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

Ter uma lista documentada dos requisitos de rede da workload, incluindo latência, tamanho de pacotes, regras de roteamento, protocolos e padrões de tráfego compatíveis. Analise as soluções de redes disponíveis e identifique os serviços que atendem às características de 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.

Etapas da implementação:

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

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

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

Recursos

Documentos relacionados:

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

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (Conectividade com a AWS e arquiteturas de rede híbrida da AWS)
- Optimizing Network Performance for Amazon EC2 Instances (Otimização da performance da rede para instâncias do Amazon EC2)
- Improve Global Network Performance for Applications (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 e soluções de segurança escaláveis
- Workshops de redes da AWS

PERF04-BP02 Avaliar os recursos de redes disponíveis

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

Antipadrões comuns:

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

Benefícios de estabelecer esta prática recomendada: avaliar todos os recursos e opções de serviços pode aumentar a performance da workload, reduzir o custo da infraestrutura, diminuir o esforço necessário para manter sua workload e aumentar sua postura geral de segurança. É possível utilizar 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 para implementação

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

Analise quais opções de configuração de rede estão disponíveis e como elas poderiam afetar a workload. A otimização do desempenho depende da compreensão de como essas opções interagem com sua arquitetura e do impacto que elas terão no desempenho medido e na experiência do usuário.

Etapas da implementação

- Crie uma lista de componentes da workload.

  - Considere o uso de Nuvem AWS WAN para criar, gerenciar e monitorar a rede da sua organização ao criar uma rede global unificada.
  - Monitore suas redes globais e centrais com métricas do Amazon CloudWatch Logs. Utilize o Amazon CloudWatch RUM, que fornece insights para ajudar a identificar, entender e aprimorar a experiência digital dos usuários.
  - Visualize a latência agregada da rede entre Regiões da AWS e Zonas de Disponibilidade, bem como dentro de cada Zona de Disponibilidade, usando AWS Network Manager para obter informações sobre como o desempenho da sua aplicação se relaciona com o desempenho da rede da AWS subjacente.
  - 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.
- 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.
- Se a workload for nova, realize testes de carga para identificar gargalos de performance.
- Para os gargalos de performance que identificar, analise as opções de configuração para suas soluções a fim de identificar oportunidades de melhoria da performance. Confira as seguintes principais opções e recursos de rede:

Oportunidade de melhoria	Solução
Caminho ou rotas de rede	Use o Network Access Analyzer para identific ar caminhos ou rotas.
Protocolos de rede	Consulte PERF04-BP05 Escolher os protocolo s de rede para melhorar o desempenho
Topologia de rede	Avalie suas concessões de operação e performance entre Emparelhamento de VPC e AWS Transit Gateway ao conectar várias contas. O AWS Transit Gateway simplifica a forma como você interconecta todas as suas

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

Oportunidade de melhoria	Solução
Recursos do atributo de armazenamento	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. 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.

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

Recursos

Documentos relacionados:

- 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 Adaptador de Rede Elástica (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

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (Conectividade com a AWS e arquiteturas de rede híbrida da AWS)
- Optimizing Network Performance for Amazon EC2 Instances (Otimização da performance da rede para instâncias do Amazon EC2)
- AWS Global Accelerator

Exemplos relacionados:

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

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

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

Antipadrões comuns:

- Você só avalia as soluções de VPN para seus requisitos de criptografia de rede.
- Você não avalia as opções de backup ou de conectividade redundante.
- Você não identifica todos os requisitos da workload (necessidades de criptografia, protocolo, largura de banda e tráfego).

Benefícios de estabelecer esta prática recomendada: Selecionar e configurar soluções de conectividade apropriadas aumentará a confiabilidade da workload e maximizará a performance. A identificação dos requisitos da workload, o planejamento antecipado e a avaliação das soluções híbridas podem minimizar alterações dispendiosas da rede física e despesas operacionais, e aumentará seu time-to-value.

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

Orientação para implementação

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

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

Etapas da implementação

1. Calcule os requisitos de largura de banda e latência de suas aplicações existentes.
   a. Para workloads existentes que estão sendo migradas para a AWS, utilize os dados de seus sistemas de monitoramento de rede internos.
   b. Para workloads novas ou existentes para as quais não há dados de monitoramento, consulte os proprietários do produto para determinar métricas de performance adequadas e fornecer uma experiência do usuário satisfatória.
2. Escolha uma conexão dedicada ou VPN como sua opção de conectividade. Com base em todos os requisitos da workload (necessidades de criptografia, largura de banda e tráfego), é possível

escolher o AWS Direct Connect ou a AWS VPN (ou ambos). O diagrama a seguir ajudará você a escolher o tipo de conexão apropriada.

a. O AWS Direct Connect fornece conectividade dedicada ao ambiente da AWS, de 50 Mbps a 100 Gbps, usando conexões dedicadas ou conexões hospedadas. Isso permite que você tenha latência gerenciada e controlada, além de largura de banda provisionada para que a workload possa se conectar de forma eficiente com outros ambientes. Com os parceiros do AWS Direct Connect, é possível ter conectividade completa para vários ambientes, fornecendo uma rede estendida com performance consistente. A AWS oferece escalabilidade da largura de banda da conexão direta usando o grupo de agregação nativo (LAG) de 100 Gbps ou o BGP equal-cost multipath (ECMP).

b. A AWS Site-to-Site VPN fornece um serviço de VPN gerenciada compatível com o protocolo de segurança da internet (IPsec). Quando uma conexão VPN é criada, cada conexão VPN inclui dois túneis para alta disponibilidade.

3. Siga a documentação da AWS para escolher uma opção de conectividade apropriada:

a. Se você decidir usar o AWS Direct Connect, selecione a largura de banda apropriada para sua conectividade.

b. Se você usar uma AWS Site-to-Site VPN em vários locais para se conectar a uma Região da AWS, use uma conexão de Site-to-Site VPN acelerada para melhorar a performance da rede.

c. Se o design da sua rede consistir em uma conexão VPN IPSec no AWS Direct Connect, considere o uso de VPN de IP privado para melhorar a segurança e conseguir segmentação. A AWS Site-to-Site Private IP VPN é implantada sobre a interface virtual de trânsito (VIF).

d. O AWS Direct Connect SiteLink permite criar conexões redundantes e de baixa latência entre seus datacenters em todo o mundo, enviando dados pelo caminho mais rápido entre os locais do AWS Direct Connect, contornando Regiões da AWS.

4. Valide sua configuração de conectividade antes de implantá-la na produção. Execute testes de segurança e performance para garantir que ela atenda aos requisitos de largura de banda, confiabilidade, latência e conformidade.

5. Monitore regularmente a performance e o uso da conectividade e otimize, se necessário.

Fluxograma de desempenho determinístico

Recursos

Documentos relacionados:

- Network Load Balancer
- Produtos de redes com a AWS
- AWS Transit Gateway
- Fazer a transição para o encaminhamento por latência no Amazon Route 53
- Endpoints 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)
- AWS Direct Connect
- Client VPN

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (Conectividade com a AWS e arquiteturas de rede híbrida da AWS)
- Optimizing Network Performance for Amazon EC2 Instances (Otimização da performance da rede para instâncias do Amazon EC2)
- AWS Global Accelerator
- AWS Direct Connect
- AWS Transit Gateway Connect
- Soluções de VPN
- Segurança com as soluções de VPN

Exemplos relacionados:

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

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

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

Antipadrões comuns:

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

Benefícios de estabelecer esta prática recomendada: Um balanceador de carga lida com a carga variável do tráfego da sua aplicação em uma única Zona de Disponibilidade ou em várias Zonas de Disponibilidade e permite alta disponibilidade, ajuste de escala automático e melhor utilização de sua workload.

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

Orientação para implementação

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

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

A AWS fornece vários modelos para que suas aplicações usem o balanceamento de carga. O Application Load Balancer é o mais adequado para balanceamento de carga de tráfego HTTP e HTTPS, e oferece roteamento avançado de solicitação direcionado para a entrega de arquiteturas de aplicações modernas, inclusive microsserviços e contêineres.

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

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

Depois de escolher o balanceador de carga certo, você pode começar a utilizar seus recursos para reduzir a quantidade de esforço que seu back-end precisa fazer para atender o tráfego.

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

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

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

Os requisitos de latência da workload devem ser considerados ao definir a arquitetura. Como exemplo, se você tiver uma aplicação sensível à latência, poderá decidir usar o Network Load Balancer, que oferece latências extremamente baixas. Como alternativa, você pode decidir aproximar a workload dos clientes utilizando o Application Load Balancer em zonas locais da AWS ou mesmo o AWS Outposts.

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

Use o Auto Scaling integrado ao balanceador de carga. Um dos principais aspectos de um sistema com desempenho eficiente está relacionado ao dimensionamento correto dos recursos de back-end. Para fazer isso, é possível utilizar as integrações do balanceador de carga para os recursos de destino de back-end. Ao usar a integração do balanceador de carga com os grupos do Auto Scaling, os destinos serão adicionados ou removidos do balanceador de carga conforme exigido em resposta

ao tráfego recebido. Os balanceadores de carga também podem se integrar com o Amazon ECS e o Amazon EKS para workloads em contêineres.

- Amazon ECS: balanceamento de carga do serviço
- Balanceamento de carga da aplicação no Amazon EKS
- Balanceamento de carga da rede no Amazon EKS

Etapas da implementação

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

    - Round robin: use quando as solicitações e os destinos forem semelhantes, ou se você precisar distribuir as solicitações igualmente entre os destinos.
- Considere isolamento por zona ou entre zonas.
    - Desative a opção entre zonas (isolamento por zona) para melhorias de latência e domínios com falha de zona. Ele está desativado por padrão no NLB e no ALB. Você pode desativá-lo por grupo-alvo.
    - Ative a opção entre zonas para maior disponibilidade e flexibilidade. Por padrão, a opção entre zonas está ativada para o ALB. No NLB, você pode ativá-la por grupo-alvo..
- Ative as manutenções de funcionamento de HTTP para as workloads HTTP (apenas ALB). Com esse recurso, o balanceador de carga pode reutilizar as conexões de back-end até expirar o tempo limite da manutenção de funcionamento, melhorando a solicitação HTTP e o tempo de resposta, além de reduzir a utilização de recursos nos destinos de back-end. Para obter detalhes sobre como fazer isso para Apache e Nginx, consulte Quais são as configurações ideais para usar o Apache ou o NGINX como servidor de back-end para o ELB?
- Ative o monitoramento do balanceador de carga.
    - Ative os logs de acesso para o Application Load Balancer e o Network Load Balancer.
    - Os principais campos a considerar para o ALB são `request_processing_time`, o `request_processing_timee` o `response_processing_time`.
    - Os principais campos a considerar para o NLB são `connection_time` e o `tls_handshake_time`.
    - Esteja pronto para consultar os logs quando precisar deles. Você pode usar o Amazon Athena para consultar tanto os logs do ALB e os logs do NLB.
    - Crie alarmes para métricas relacionadas ao desempenho, como `TargetResponseTime` para o ALB.

Recursos

Documentos relacionados:

- Comparação de produtos do ELB
- AWS Global Infrastructure (Infraestrutura global da AWS)
- Improving Performance and Reducing Cost Using Availability Zone Affinity (Melhorar o desempenho e reduzir os custos usando a afinidade de zona de disponibilidade)

- Step by step for Log Analysis with Amazon Athena (Passo a passo para a análise de logs com o Amazon Athena)
- Querying Application Load Balancer logs (Consulta de logs do Application Load Balancer)
- Monitor your Application Load Balancers (Monitore o Application Load Balancers)
- Monitor your Network Load Balancer (Monitore o Network Load Balancer)
- Use Elastic Load Balancing to distribute traffic across the instances in your Auto Scaling group (Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling)

Vídeos relacionados:

- AWS re:Invent 2018: Elastic Load Balancing: Deep Dive and Best Practices (AWS re:Invent 2018: Elastic Load Balancing: aprofundamento e práticas recomendadas)
- AWS re:Invent 2021 - How to choose the right load balancer for your AWS workloads (AWS re:Invent 2021: como escolher o balanceador de carga certo para suas workloads da AWS)
- AWS re:Inforce 2022 - How to use Elastic Load Balancing to enhance your security posture at scale (AWS re:Inforce 2022: como usar o Elastic Load Balancing para melhorar seu procedimento de segurança em escala)
- AWS re:Invent 2019: Get the most from Elastic Load Balancing for different workloads (AWS re:Invent 2019: aproveite ao máximo o Elastic Load Balancing para diferentes workloads)

Exemplos relacionados:

- CDK and AWS CloudFormation samples for Log Analysis with Amazon Athena (Exemplos de CDK e AWS CloudFormation para análise de log com o Amazon Athena)

PERF04-BP05 Escolher os protocolos de rede para melhorar o desempenho

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

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

Antipadrões comuns:

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

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

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

Orientação para implementação

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

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

Quando considerar o uso do TCP

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

Quando considerar o uso do UDP

O UDP é um protocolo sem conexão e, portanto, é adequado para aplicações que precisam de uma transmissão rápida e eficiente, como log, monitoramento e dados de VoIP. Além disso,

considere usar o UDP se você tiver componentes da workload que respondam a pequenas consultas de grandes números de clientes para garantir um desempenho ideal da workload. O Datagram Transport Layer Security (DTLS) é o equivalente UDP do Transport Layer Security (TLS). Ao usar DTLS com UDP, a sobrecarga vem da criptografia e descriptografia de dados, já que o processo de handshake é simplificado. O DTLS também adiciona uma pequena quantidade de sobrecarga aos pacotes de UDP, já que inclui campos adicionais para indicar os parâmetros de segurança e detectar violações.

Quando considerar o uso do SRD

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

Etapas da implementação

1. Use o AWS Global Accelerator e o AWS Transfer Family para melhorar o throughput de suas aplicações de transferência de arquivos online. O serviço AWS Global Accelerator ajuda você a obter baixa latência entre os dispositivos cliente e a workload na AWS. Com o AWS Transfer Family, é possível usar protocolos baseados em TCP, como SFTP (Protocolo de transferência de arquivos de Secure Shell) e FTPS (Protocolo de transferência de arquivos por SSL), para escalar e gerenciar com segurança as transferências de arquivos para os serviços de armazenamento da AWS.
2. Use a latência de rede para determinar se o TCP é adequado para comunicação entre os componentes da workload. Se a latência de rede entre a aplicação cliente e o servidor for alta, o handshake de três vias do TCP pode levar um tempo, afetando, assim, a capacidade de resposta da aplicação. Métricas como tempo até o primeiro byte (TTFB) e tempo de ida e volta (RTT) podem ser usadas para medir a latência da rede. Se sua workload fornece conteúdo dinâmico aos usuários, considere usar o Amazon CloudFront, que estabelece uma conexão persistente com cada origem de conteúdo dinâmico para remover o tempo de configuração da conexão que, de outra forma, diminuiria a velocidade de cada solicitação do cliente.
3. Usar TLS com TCP ou UDP pode resultar em maior latência e menor throughput para a workload devido ao impacto da criptografia e descriptografia. Para essas workloads, considere o descarregamento de SSL/TLS no Elastic Load Balancing para melhorar o desempenho

da workload, permitindo que o balanceador de carga lide com o processo de criptografia e descriptografia de SSL/TLS em vez de deixar que as instâncias de back-end façam isso. Isso pode ajudar a reduzir a utilização da CPU nas instâncias de back-end, o que pode melhorar o desempenho e aumentar a capacidade.

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

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 Adaptador de Rede Elástica (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

Vídeos relacionados:

- Connectivity to AWS and hybrid AWS network architectures (Conectividade com a AWS e arquiteturas de rede híbrida da AWS)
- Optimizing Network Performance for Amazon EC2 Instances (Otimização da performance da rede para instâncias do Amazon EC2)

Exemplos relacionados:

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

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

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

Antipadrões comuns:

- Você consolida todos os recursos da workload em uma única localização geográfica.
- Você escolhe a Região mais próxima ao seu local, mas não ao usuário final da workload.

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

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

Orientação para implementação

Recursos, como instâncias do Amazon EC2, são colocados em zonas de disponibilidade em Regiões da AWS, zonas locais da AWS, AWS Outpostsou AWS Wavelength . A escolha desse local influencia o throughput e a latência da rede de determinado local do usuário. Serviços de borda, como Amazon

CloudFront e o AWS Global Accelerator também podem ser usados para melhorar o desempenho da rede, seja armazenando o conteúdo em cache nos locais da borda ou oferecendo aos usuários um ótimo caminho para a workload por meio da rede global da AWS.

O Amazon EC2 oferece grupos de posicionamento para redes. Um grupo de posicionamento é um agrupamento lógico de instâncias para diminuir a latência. O uso de grupos de posicionamento com tipos de instância compatíveis e um Adaptador de Rede Elástica (ENA) permite que as workloads participem de uma rede de baixa latência, com oscilação reduzida e de 25 Gbps. Recomenda-se o uso de grupos de posicionamento para workloads que se beneficiam de baixa latência de rede, alto throughput de rede ou ambos.

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

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

Etapas da implementação

- Capture informações sobre o tráfego IP que entra e sai das interfaces de rede.
  - Registro em log do tráfego IP usando logs de fluxo de VPC
  - Como o endereço IP do cliente é preservado no AWS Global Accelerator
- Analise os padrões de acesso à rede em sua workload para identificar como os usuários utilizam sua aplicação.
  - Use ferramentas de monitoramento, como Amazon CloudWatch e o AWS CloudTrail, para coletar dados sobre as atividades da rede.
  - Analise os dados para identificar o padrão de acesso à rede.
- Selecione as Regiões para implantação da workload com base nos seguintes elementos fundamentais:
  - 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 seus usuários: para aplicações voltadas ao usuário, escolha uma Região (ou Regiões) próxima dos clientes de sua workload.
  - Outras restrições: leve em conta restrições, como custo e conformidade, conforme explicado em O que considerar ao selecionar uma região para suas workloads.
- Use zonas locais da AWS para executar workloads como renderização de vídeo. As zonas locais permitem que você se beneficie de ter recursos de computação e armazenamento mais próximos dos usuários finais.
- Use AWS Outposts para workloads que precisam permanecer on-premises e onde você deseja que essa workload seja executada ininterruptamente com o restante de suas workloads na AWS.
- Aplicações, como streaming de vídeo ao vivo em alta resolução, áudio de alta fidelidade ou realidade aumentada/realidade virtual (RA/RV), exigem latência ultrabaixa para dispositivos 5G. Para tais aplicações, considere o AWS WavelengthO AWS Wavelength incorpora serviços de armazenamento e computação da AWS em redes 5G, fornecendo a infraestrutura móvel de computação de borda para desenvolver, implantar e escalar aplicações de latência ultrabaixa.
- Use armazenamento em cache local ou soluções de armazenamento em cache da AWS para ativos usados com frequência a fim de aumentar a performance, reduzir a movimentação de dados e reduzir o impacto ambiental.

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

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

Serviço	Quando usar
Lambda@edge	Use para operações com uso intenso de computação que são iniciadas quando objetos não estão no cache.
Funções do Amazon CloudFront	Use para casos de uso simples, como solicitações HTTP(s) ou manipulações de resposta, que podem ser iniciadas por funções de curta duração.
AWS IoT Greengrass	Use para executar computação local, mensagens e armazenamento de dados em cache para dispositivos conectados.

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

Recursos

Práticas recomendadas relacionadas:

- COST07-BP02 Implementar regiões com base nos custos
- COST08-BP03 Implementar serviços para reduzir custos de transferência de dados
- REL10-BP01 Implantar a workload em vários locais
- REL10-BP02 Escolher os locais apropriados para sua implantação de vários locais

- SUS01-BP01 Escolher a região com base nos requisitos empresariais e nas metas de sustentabilidade
- SUS02-BP04 Otimizar o posicionamento geográfico das workloads com base nos respectivos requisitos de rede
- SUS04-BP07 Minimizar a movimentação de dados entre redes

Documentos relacionados:

- AWS Global Infrastructure (Infraestrutura global da AWS)
- AWS Local Zones and AWS Outposts, choosing the right technology for your edge workload (Zonas locais da AWS e AWS Outposts: como escolher a tecnologia certa para sua workload de borda)
- Grupos de posicionamento
- zonas locais da AWS
- AWS Outposts
- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Vídeos relacionados:

- AWS Local Zones Explainer Video (Vídeo de explicação de zonas locais da AWS)
- AWS Outposts: Overview and How it Works (AWS Outposts: visão geral e como funciona
- AWS re:Invent 2021 - AWS Outposts: Bringing the AWS experience on premises (AWS re:Invent 2021 - AWS Outposts: como trazer a experiência da AWS para ambientes on-premises)
- AWS re:Invent 2020: AWS Wavelength: Run apps with ultra-low latency at 5G edge (AWS re:Invent 2020: AWS Wavelength: execute aplicativos com latência ultrabaixa na borda 5G)
- AWS re:Invent 2022 - AWS Local Zones: Building applications for a distributed edge (AWS re:Invent 2022: zonas locais da AWS: como criar aplicações para uma borda distribuída)

- AWS re:Invent 2021 - Building low-latency websites with Amazon CloudFront (AWS re:Invent 2021: criação de sites de baixa latência com o Amazon CloudFront)
- AWS re:Invent 2022 - Improve performance and availability with AWS Global Accelerator (AWS re:Invent 2022: melhore a performance e a disponibilidade com o AWS Global Accelerator)
- AWS re:Invent 2022 - Build your global wide area network using AWS (AWS re:Invent 2022: crie sua rede de longa distância usando a AWS)
- AWS re:Invent 2020: Global traffic management with Amazon Route 53 (AWS re:Invent 2020: gerenciamento de tráfego global com o Amazon Route 53)

Exemplos relacionados:

- Workshop do AWS Global Accelerator
- Handling Rewrites and Redirects using Edge Functions (Lidar com reescritas e redirecionamentos usando funções da borda)

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

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

Antipadrões comuns:

- Você pressupõe que todos os problemas relacionados à performance são relacionados à aplicação.
- Você só testa a performance da rede a partir de um local próximo ao local em que implantou a carga de trabalho.
- Você usa configurações-padrão para todos os serviços de rede.
- Você provisiona em excesso recursos de rede para fornecer capacidade suficiente.

Benefícios de estabelecer esta prática recomendada: coletar as métricas necessárias da rede da AWS e implementar ferramentas de monitoramento de rede permite entender o desempenho da rede e otimizar as configurações dela.

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

Orientação para implementação

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

Etapas da implementação

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

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

Ferramenta da AWS	Onde usar
	o o domínio ou subdomínio solicitado ou os locais da borda do Route 53 que responderam às consultas ao DNS.
Reachability Analyzer	O Reachability Analyzer ajuda você a analisar e depurar a capacidade de alcance da rede. O Reachability Analyzer é uma ferramenta de análise de configuração que permite realizar testes de conectividade entre um recurso da origem e um do destino nas VPCs. Essa ferramenta ajuda a verificar se a configura ção da rede corresponde à conectividade pretendida.
Network Access Analyzer	O Network Access Analyzer ajuda você a entender o acesso da rede aos seus recursos. É possível usar o Network Access Analyzer para especificar os requisitos de acesso à rede e identificar possíveis caminhos de rede que não atendem aos requisitos especific ados. Ao otimizar a configuração da rede correspondente, é possível entender e verificar o estado da rede e demonstrar se a rede na AWS atende aos seus requisitos de conformidade.

Ferramenta da AWS	Onde usar
Amazon CloudWatch	Use o Amazon CloudWatch e ative as métricas apropriadas para as opções de rede. Escolha a métrica de rede certa para sua workload. Por exemplo, é possível habilitar métricas para o uso do endereço de rede da VPC, o gateway NAT da VPC, o AWS Transit Gateway, o túnel da VPN, o AWS Network Firewall, o Elastic Load Balancing e o AWS Direct Connect. Monitorar continuamente as métricas é uma prática recomendada para observar e entender o status e o uso da rede, o que ajuda a otimizar a configuração da rede com base em suas observações.
AWS Network Manager	Usando o AWS Network Manager, você pode monitorar o desempenho histórico e em tempo real da rede global da AWS para fins operacionais e de planejamento. O Network Manager fornece latência de rede agregada entre as Regiões da AWS e Zonas de Disponibilidade e dentro de cada Zona de Disponibilidade, permitindo que você entenda melhor como o desempenho da sua aplicação se relaciona com o desempenho da rede da AWS subjacente.
Amazon CloudWatch RUM	Use o Amazon CloudWatch RUM para coletar as métricas que fornecem os insights que ajudam a identificar, entender e melhorar a experiência do usuário.

- Identifique os principais interlocutores e os padrões de tráfego de aplicações usando VPC e logs de fluxo do AWS Transit Gateway.

- Avalie e otimize sua arquitetura de rede atual, incluindo VPCs, sub-redes e roteamento. Como exemplo, você pode avaliar como diferentes emparelhamentos de VPC ou AWS Transit Gateway podem ajudar a melhorar a rede em sua arquitetura.
- Avalie os caminhos de roteamento em sua rede para verificar se o caminho mais curto entre os destinos é sempre usado. O Network Access Analyzer pode ajudar nessa tarefa.

Recursos

Documentos relacionados:

- Logs de fluxo da VPC
- Registro em log de consulta ao DNS público
- O que é o IPAM?
- O que é o Reachability Analyzer?
- O que é o Network Access Analyzer?
- Métricas do CloudWatch para suas VPCs
- Otimize o desempenho e reduza os custos de análise da rede com os logs de fluxo da VPC no formato Apache Parquet
- Monitoramento de suas redes globais e principais com métricas do Amazon CloudWatch
- Monitore continuamente o tráfego e os recursos da rede

Vídeos relacionados:

- Networking best practices and tips with the AWS Well-Architected Framework (Práticas recomendadas e dicas de redes com o AWS Well-Architected Framework)
- Monitoring and troubleshooting network traffic (Monitoramento e resolução de problemas de tráfego de rede)

Exemplos relacionados:

- Workshops de redes da AWS
- Monitoramento de rede da AWS

Processo e cultura

PERFORMANCE 5. Como suas práticas e cultura organizacionais contribuem para a eficiência do desempenho em sua workload?

Ao arquitetar workloads, há princípios e práticas que você pode adotar para ajudar na melhor execução de workloads de nuvem eficientes e de alto desempenho. Para adotar uma cultura que promova a eficiência do desempenho das workloads na nuvem, considere estes princípios e práticas fundamentais:

Práticas recomendadas

- PERF05-BP01 Estabeleça indicadores-chave de desempenho (KPIs) para medir a integridade e o desempenho da workload
- PERF05-BP02 Use soluções de monitoramento para entender as áreas em que o desempenho é mais crítico
- PERF05-BP03 Defina um processo para melhorar a performance da workload
- PERF05-BP04 Faça o teste de carga da workload
- PERF05-BP05 Use a automação para corrigir proativamente problemas relacionados ao desempenho
- PERF05-BP06 Mantenha a workload e os serviços atualizados
- PERF05-BP07 Analise as métricas regularmente

PERF05-BP01 Estabeleça indicadores-chave de desempenho (KPIs) para medir a integridade e o desempenho da workload

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

Antipadrões comuns:

- Você só monitora as métricas no nível do sistema para obter informações da workload e não compreende aos impactos dessas métricas nos negócios.
- Você pressupõe que os KPIs já estejam publicados e compartilhados como dados de métricas comuns.
- Você não define um KPI quantitativo e mensurável.

- Você não alinha os KPIs às metas ou estratégias empresariais.

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

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

Orientação para implementação

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

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

Etapas da implementação

1. Identifique e documente as principais partes interessadas da empresa.
2. Trabalhe com essas partes interessadas para definir e documentar os objetivos da workload.
3. Analise as práticas recomendadas do setor para identificar KPIs relevantes alinhados aos objetivos da workload.

4. Use as práticas recomendadas do setor e os objetivos da workload para definir metas de KPI da workload. Use essas informações para definir limites de KPI no nível de gravidade ou de alarme.
5. Identifique e documente o risco e o impacto no caso de um KPI não ser atendido.
6. Identifique e documente métricas que podem ajudar a estabelecer os KPIs.
7. Use ferramentas de monitoramento, como Amazon CloudWatch ou AWS Config para coletar métricas e medir KPIs.
8. Use painéis para visualizar e comunicar os KPIs com as partes interessadas.
9. Revise e analise regularmente as métricas para identificar áreas da workload que precisam ser aprimoradas.
10. Revise os KPIs quando as metas empresariais ou a performance da workload mudarem.

Recursos

Documentos relacionados:

- Documentação da CloudWatch
- AWS Partners de monitoramento, registro em log e performance
- Documentação do X-Ray
- Using Amazon CloudWatch dashboards
- Amazon QuickSight KPIs

Vídeos relacionados:

- AWS re:Invent 2019: Scaling up to your first 10 million users
- Cut through the chaos: Gain operational visibility and insight
- Build a monitoring plan

Exemplos relacionados:

- Creating a dashboard with Amazon QuickSight

PERF05-BP02 Use soluções de monitoramento para entender as áreas em que o desempenho é mais crítico

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.

Antipadrões comuns:

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

Benefícios de 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.

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. Depois de investigar essas áreas, veja qual solução você pode implantar para eliminar esses problemas de desempenho.

Etapas da implementação

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

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

2. Realize testes para gerar métricas, identificar padrões de tráfego, gargalos e áreas de desempenho críticas. Aqui estão alguns exemplos de como realizar testes:
    - Configure o Canários sintéticos do CloudWatch para imitar programaticamente as atividades do usuário baseadas no navegador usando trabalhos cron do Linux ou expressões de taxa para gerar métricas consistentes ao longo do tempo.
    - Use o 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. Por exemplo, você pode usar o CloudWatch Evidently para testar novas melhorias e impactos na performance da workload.

Recursos

Documentos relacionados:

- Amazon Builders' Library
- Documentação do X-Ray
- Amazon CloudWatch RUM
- Amazon DevOps Guru

Vídeos relacionados:

- The Amazon Builders' Library: 25 years of Amazon operational excellence
- Visual Monitoring of Applications with 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

PERF05-BP03 Defina 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.

Antipadrões comuns:

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

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

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

Orientação para implementação

A performance 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.

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

Etapas da implementação

- Identifique seus KPIs de performance da workload conforme descrito em PERF05-BP01 Estabeleça indicadores-chave de desempenho (KPIs) para medir a integridade e o desempenho da workload para basear sua workload.
- Use Ferramentas de observabilidade da AWS para coletar métricas de performance e medir KPIs.
- Faça uma análise aprofundada para identificar as áreas (como configuração e código da aplicação) na workload que estão com baixa performance, conforme descrito em PERF05-BP02 Use soluções de monitoramento para entender as áreas em que o desempenho é mais crítico.
- Use suas ferramentas de análise e desempenho para identificar a estratégia de otimização de desempenho.

- Use ambientes de sandbox ou de pré-produção para validar a eficácia da estratégia.
- Implemente as mudanças na produção e monitore constantemente o desempenho da workload.
- Documente as melhorias e comunique isso às partes interessadas.

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

PERF05-BP04 Faça o teste de carga da workload

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

Antipadrões comuns:

- Você realiza um teste de carga de peças individuais da workload, mas não toda a workload.
- 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.

- Você realiza testes de carga sem consultar a política de testes do Amazon EC2 e enviar um formulário de envio de eventos simulados. Isso faz com que o teste não seja executado, pois parece um evento de negação de serviço.

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

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

Orientação para implementação

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

Etapas da implementação

- Configure o ambiente de teste com base no ambiente de produção. É possível usar os serviços da AWS para executar ambientes em escala de produção para testar a arquitetura.
- Escolha e configure a ferramenta de teste de carga adequada à workload.
- Defina os cenários e parâmetros do teste de carga (como duração do teste e número de usuários).
- Execute cenários de teste em grande escala. Aproveite a Nuvem AWS para testar a workload e descobrir se há uma falha na escala ou se ela está com a escala reduzida horizontalmente de maneira não linear. Por exemplo, use instâncias spot para gerar cargas a um baixo custo e descobrir gargalos antes que eles ocorram em produção.
- Monitore e registre métricas de desempenho (como throughput e tempo de resposta). O Amazon CloudWatch pode coletar métricas entre os recursos em sua arquitetura. Você também pode coletar e publicar métricas personalizadas para descobrir métricas de negócio ou derivadas.
- Analise os resultados para identificar gargalos de desempenho e áreas para melhorias.
- Documente e relate o processo e os resultados do teste de carga.

Recursos

Documentos relacionados:

- AWS CloudFormation
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Testes de carga distribuída na AWS

Vídeos relacionados:

- Solving with AWS Solutions: Distributed Load Testing

- Optimize applications through Amazon CloudWatch RUM
- Demo of Amazon CloudWatch Synthetics (Demonstração do Amazon CloudWatch Synthetics)

Exemplos relacionados:

- Testes de carga distribuída na AWS

PERF05-BP05 Use a automação para corrigir proativamente problemas relacionados ao desempenho

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

Antipadrões comuns:

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

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

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

Orientação para implementação

Sempre que possível, use alarmes para desencadear ações automatizadas visando corrigir problemas. Se a resposta automatizada não for possível, encaminhe o alarme para aqueles capazes de responder. Por exemplo, você pode ter um sistema capaz de prever os valores de indicadores-chave de desempenho (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.

Etapas da implementação

- Identifique e compreenda o problema de desempenho que pode ser corrigido automaticamente. Use soluções de monitoramento da AWS, como o Amazon CloudWatch ou o AWS X-Ray, para ajudar você a entender melhor a causa raiz do problema.
- Crie um plano e um processo de correção detalhados que possam ser usados para corrigir automaticamente o problema.
- Configure o gatilho para iniciar automaticamente o processo de correção. Por exemplo, você pode definir um acionador para reiniciar automaticamente uma instância quando ela atinge determinado limite de utilização da CPU.
- Use serviços e tecnologias da AWS para automatizar o processo de correção. Por exemplo: AWS Systems Manager Automation fornece uma maneira segura e escalável de automatizar o processo de correção.
- Teste o processo de correção automatizado em um ambiente de pré-produção.
- Após o teste, implemente o processo de correção no ambiente de produção e monitore constantemente para identificar áreas de melhoria.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Parceiros da AWS Partner Network de monitoramento, registro em log e performance
- Documentação do X-Ray

- Using Alarms and Alarm Actions in CloudWatch

Vídeos relacionados:

- Intelligently automating cloud operations (Automatizar de forma inteligente as operações na nuvem)
- Setting up controls at scale in your AWS environment
- Automating patch management and compliance using AWS
- How Amazon uses better metrics for improved website performance (Como a Amazon usa métricas melhores para aprimorar o desempenho do site)

Exemplos relacionados:

- CloudWatch Logs Customize Alarms

PERF05-BP06 Mantenha a workload e os serviços atualizados

Atualize-se com relação aos novos serviços e atributos de nuvem para adotar recursos eficientes, remover problemas e melhorar a eficiência geral do desempenho da workload.

Antipadrões comuns:

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

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

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

Orientação para implementação

Avalie maneiras de melhorar o desempenho conforme são disponibilizados novos serviços, padrões de design e atributos 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, crie uma prova de conceito que use novas tecnologias ou consulte 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.

Etapas da implementação

- Fazer o inventário de software e arquitetura da workload e identificar os componentes que precisam ser atualizados.
- Identifique novidades e atualize fontes relacionadas aos componentes da workload. Por exemplo, você pode se inscrever no What's New at AWS para os produtos que correspondem ao componente da workload. Você pode assinar o feed RSS ou gerenciar suas assinaturas de e-mail.
- Defina um cronograma para avaliar novos serviços e atributos para a workload.
  - 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. Aproveite a agilidade da nuvem para testar rapidamente como novos atributos podem melhorar a workload com o intuito de obter eficiências de performance.
- Use automação no processo de atualização para reduzir o nível de esforço para implantar novos recursos e limitar erros causados por processos manuais.
  - Você pode usar o CI/CD para atualizar automaticamente AMIs, imagens de contêiner e outros artefatos relacionados à aplicação de nuvem.
  - Você pode usar ferramentas, como 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.
- Documente seu processo para avaliar atualizações e novos serviços. Forneça aos proprietários o tempo e o espaço necessários para pesquisar, testar, experimentar e validar atualizações e novos serviços. Consulte novamente os KPIs e requisitos empresariais documentados para ajudar a priorizar qual atualização trará um impacto positivo à empresa.

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:

- Laboratórios do Well-Architected: Gerenciamento de inventário e patches
- Laboratório: AWS Systems Manager

PERF05-BP07 Analise as métricas regularmente

Como parte da 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 poderiam ajudar a identificar, resolver ou prevenir problemas se estivessem sendo acompanhadas.

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

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

Orientação para implementação

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

qualidade das métricas coletadas, de modo que você possa prevenir ou resolver incidentes futuros mais rapidamente.

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

Etapas da implementação

1. Defina métricas essenciais de desempenho a serem monitoradas que estejam alinhadas ao seu objetivo de workload.
2. Defina uma linha de base e um valor desejável para cada métrica.
3. Defina uma frequência (como semanal ou mensal) para revisar as métricas essenciais.
4. Durante cada revisão, avalie as tendências e o desvio dos valores base. Procure gargalos ou anomalias de desempenho.
5. Para os problemas identificados, realize uma análise aprofundada da causa raiz para entender o principal motivo do problema.
6. Documente as descobertas e use estratégias para lidar com os problemas e gargalos identificados.
7. Avalie e melhore constantemente o processo de revisão de métricas.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Collect metrics and logs from Amazon EC2 Instances and on-premises servers with the CloudWatch Agent
- Parceiros da AWS Partner Network de monitoramento, registro em log e performance
- Documentação do X-Ray

Vídeos relacionados:

- Setting up controls at scale in your AWS environment

- How Amazon uses better metrics for improved website performance (Como a Amazon usa métricas melhores para aprimorar o desempenho do site)

Exemplos relacionados:

- Creating a dashboard with Amazon QuickSight
- Level 100: Monitoring with CloudWatch Dashboards

Otimização de custos

O pilar Otimização de custos inclui a capacidade de executar sistemas para proporcionar valor comercial pelo menor preço. Você pode encontrar orientações prescritivas sobre implementação no whitepaper sobre o pilar de otimização de custos.

Áreas de práticas recomendadas

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

Pratique o gerenciamento financeiro na nuvem

Pergunta

- CUSTOS 1. Como implementar o gerenciamento financeiro na nuvem?

CUSTOS 1. Como implementar o gerenciamento financeiro na nuvem?

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

Práticas recomendadas

- COST01-BP01 Estabelecer a propriedade da otimização de custos
- COST01-BP02 Estabelecer uma parceria entre finanças e tecnologia

- COST01-BP03 Estabelecer 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-BP08 Criar uma cultura com reconhecimento de custos
- COST01-BP09 Quantificar o valor comercial proveniente da otimização de custos

COST01-BP01 Estabelecer a propriedade da otimização de custos

Crie uma equipe (Escritório de Negócios na Nuvem, Centro de Excelência da Nuvem ou equipe FinOps) responsável por estabelecer e manter o reconhecimento de custos em toda a organização. O responsável pela otimização de custos pode ser uma pessoa ou uma equipe (requer pessoal das equipes de finanças, tecnologia e negócios) que conheça toda a organização e as finanças da nuvem.

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

Orientação para implementação

Esta é a introdução de uma função ou 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, essa 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.

Essa função (individual ou em 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 menor de tempo em atividades de gerenciamento e otimização de custos em comparação com uma função de tempo integral de uma empresa maior.

A função exige uma abordagem multidisciplinar, com recursos de gerenciamento de projetos, ciência de dados, análise financeira e desenvolvimento de software ou infraestrutura. Ela pode melhorar a eficiência da workload realizando otimizações de custos em três propriedades diferentes:

- Centralizado: por meio de equipes designadas, como a equipe FinOps, a equipe de Gerenciamento Financeiro na Nuvem (CFM), o Escritório de Negócios na Nuvem (CBO) ou o Centro de Excelência da Nuvem (CCoE), os clientes podem projetar e implementar mecanismos de governança e promover as práticas recomendadas em toda a empresa.
- Descentralizado: as equipes de tecnologia são convencidas a realizar otimizações de custos.
- Híbrido: combinação de equipes centralizadas e descentralizadas podem trabalhar em conjunto para realizar otimizações de custo.

A função pode ser medida ao comparar a sua capacidade de realização e entrega com as metas de otimização de custos (por exemplo, métricas de eficiência da workload).

Você deve garantir que haja patrocínio executivo para essa função, o que é um fator de sucesso fundamental. O patrocinador é considerado defensor do consumo de nuvem econômico e oferece suporte ao encaminhamento para a equipe a fim de garantir que as atividades de otimização de custos sejam tratadas de acordo com o nível de prioridade definido pela organização. Caso contrário, a orientação pode ser ignorada e as oportunidades de redução de custo não serão priorizadas. Juntos, o patrocinador e a equipe ajudam a organização a consumir a nuvem com eficiência e agregar valor comercial.

Se você tem um plano de suporte Business, Enterprise-On-Ramp ou Enterprise e precisa de ajuda para formar essa equipe ou função, entre em contato com especialistas em Cloud Financial Management (CFM) por meio de sua equipe de contas.

Etapas da implementação

- Defina os membros principais: todas as partes relevantes da organização devem contribuir e ter interesse pelo gerenciamento de custos. As equipes comuns dentro das organizações geralmente incluem: finanças, proprietários de aplicações ou produtos, gerenciamento e equipes técnicas (DevOps). Alguns são contratados em tempo integral (financeiro ou técnico), enquanto outros são contratados periodicamente, conforme necessário. Pessoas ou equipes encarregadas de executar o CFM precisam dos seguintes conjuntos de habilidades:
  - Desenvolvimento de software: no caso em que scripts e automação estão sendo criados.

  - Engenharia de infraestrutura: para implantar scripts, automatizar processos e entender como os serviços e os recursos são provisionados.
  - Perspicácia operacional: o intuito do CFM é permitir a operação eficiente na nuvem ao medir, monitorar, planejar e escalar o 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.

  Você pode definir indicadores-chave de performance (KPIs) baseados em valor ou custo. 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 motivadores de valor empresarial e ajudam a racionalizar mudanças em 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, empresarial 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. Depois, as principais workloads são relatadas em mais detalhes.

  Durante essas revisões regulares, você pode analisar a eficiência (custo) da workload e o resultado empresarial. 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 identificar KPIs de 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 Estabelecer 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 para 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 (ou, às vezes, aumentar) à medida que a workload ganha eficiência ou que novas workloads e atributos são implantados. As workloads podem escalar para atender 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.

Geralmente, o orçamento é preparado para um único ano e permanece fixo, exigindo adesão estrita de todos os envolvidos. Entretanto, a previsão é mais flexível, permitindo reajustes ao longo do ano e fornecendo projeções dinâmicas em um período de um, dois ou três anos. Tanto o orçamento quanto a previsão desempenham um papel fundamental para estabelecer expectativas financeiras entre várias partes interessadas em tecnologia e negócios. A precisão da previsão e implementação também impõe responsabilidade às partes interessadas que já são diretamente responsáveis pelo custo de provisionamento, e isso também pode contribuir para o reconhecimento geral de custos.

Ajuste os processos de orçamento e previsão em vigor para que se tornem mais dinâmicos por meio de um algoritmo baseado em tendências (usando custos históricos como entradas) ou de algoritmos baseados em motivadores (por exemplo, lançamentos de produtos, expansão regional ou novos ambientes para workloads), que são ideais para um ambiente de gastos dinâmico e variável, ou de uma combinação de tendências e motivadores empresariais.

Você pode usar o AWS Cost Explorer para previsões baseadas em tendências em um período futuro definido com base no gasto no passado. O mecanismo de previsão do AWS Cost Explorer segmenta os dados históricos com base em tipos de cobrança (por exemplo, instâncias reservadas) e usa uma combinação de machine learning e modelos baseados em regras com a finalidade de prever os gastos individualmente para todos os tipos de cobrança.

Identifique os motivadores empresariais que podem afetar o custo de uso e faça uma previsão para cada um deles separadamente a fim de garantir que o uso esperado seja calculado com antecedência. Alguns dos motivadores estão vinculados às equipes de TI e de produtos da organização. Outros motivadores empresariais, como eventos de marketing, promoções, fusões e aquisições, são conhecidos por seus líderes de vendas, marketing e negócios, e é importante colaborar e considerar também todos esses motivadores de demanda. Você precisa trabalhar com eles de perto para entender o impacto nos novos motivadores internos.

Depois de determinar sua previsão baseada em tendências usando o Cost Explorer ou qualquer outra ferramenta, use o AWS Pricing Calculator para estimar os custos do caso de uso da AWS e os custos futuros com base no uso esperado (tráfego, solicitações por segundo, instância do Amazon EC2 necessária). Você também pode usá-lo para planejar seus gastos, identificar oportunidades de economia e tomar decisões informadas ao usar a AWS. É importante monitorar quão precisa é essa previsão, pois os orçamentos devem ser definidos com base nesses cálculos e estimativas.

Use AWS Budgets para definir orçamentos personalizados e detalhados 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 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 pode receber um alerta quando exceder o custo ou uso, ou se houver previsão de que exceda a quantia orçada.

Use AWS Cost Anomaly Detection para evitar ou reduzir custos inesperados e aprimorar o controle sem prejudicar a inovação. O AWS Cost Anomaly Detection utiliza machine learning para

identificar gastos anômalos e causas-raiz, para que você possa agir rapidamente. Com três etapas simplesvocê pode criar seu próprio monitor contextualizado e receber alertas sempre que qualquer gasto anormal for detectado.

Conforme mencionado na seção Parceria de tecnologia e finanças do pilar de otimização de custos Well-Architected, é importante ter parceria e ritmo entre TI, departamento financeiro e outras partes interessadas para verificar se todos usam 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 para a implementação

- Analisar a previsão baseada em tendências: Use as ferramentas preferidas de previsão baseadas em tendências, como o AWS Cost Explorer e o Amazon Forecast. Analise seu custo de uso em diferentes dimensões, como serviço, conta, tags e categorias de custo. Se for necessária uma previsão avançada, importe os dados do AWS Cost and Usage Report para o Amazon Forecast (o que aplica a regressão linear como uma forma de machine learning para que seja feita a previsão).
- Analisar a previsão baseada em motivadores: identifique o impacto dos motivadores empresariais no uso da nuvem e faça uma previsão para cada um deles separadamente a fim de calcular o custo de uso esperado com antecedência. Trabalhe em estreita colaboração com proprietários de unidades de negócios e partes interessadas para entender o impacto sobre os novos motivadores e calcular as mudanças de custo esperadas para definir orçamentos precisos.
- Atualizar os processos de previsão e orçamento existentes: defina os processos orçamentários de previsão de acordo com os métodos adotados, como baseado em tendências, baseado em motivadores empresariais ou uma combinação de ambos. Os orçamentos devem ser calculados e realistas, com base nesses processos de previsão.
- Configurar alertas e notificações: use alertas do AWS Budgets e AWS Cost Anomaly Detection para receber alertas e notificações.
- 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 Cost and Usage Report
- Amazon QuickSight Forecasting
- Amazon Forecast
- AWS Budgets
- Blog de novidades da AWS

Vídeos relacionados:

- How can I use AWS Budgets to track my spending and usage
- Série de otimização de custos da AWS: AWS Budgets

Exemplos relacionados:

- Entenda e crie previsões baseadas em motivadores
- Como estabelecer e impulsionar uma cultura de previsão
- Como melhorar sua previsão de custos na nuvem
- Uso das ferramentas certas para prever custos na nuvem

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 orçamentos de nuvem e mecanismos para detectar anomalias no uso. Configure ferramentas relacionadas para alertas de custo e uso em relação a metas predefinidas e receba notificações quando algum uso exceder essas metas. Faça reuniões regulares para analisar a relação custo-benefício das workloads e promover o reconhecimento de custos.

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

Orientação para implementação

você deve informar regularmente sobre a otimização de custos e usos dentro da sua organização. Você pode implementar sessões dedicadas para discutir a relação de custo/performance ou incluir a otimização de custos em seus ciclos regulares de relatórios operacionais para as workloads. Use serviços e ferramentas para monitorar a relação de custo/performance regularmente e implementar oportunidades de redução de custos.

Visualize o custo e o uso com vários filtros e granularidade usando o AWS Cost Explorer, que fornece painéis e relatórios, como custos por serviço ou por conta, custos diários ou custos de mercado. Acompanhe o andamento do custo e do uso em relação aos orçamentos configurados com relatórios do AWS Budgets.

Uso AWS Budgets para definir orçamentos personalizados e monitorar os custos e o uso, para que possa reagir 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 programar relatórios e receber e-mails periódicos sobre o relatório de custos com o histórico de custos e uso, ou oportunidades de economia de custo. Confira nossa solução de painel de inteligência de custos (CID) integrada ao Amazon QuickSight, que oferece visibilidade avançada.

Use AWS Trusted Advisor, que oferece orientação para verificar se os recursos provisionados se alinham com as práticas recomendadas da AWS para otimização de custo.

Verifique as recomendações de Savings Plans por meio de grafos visuais em comparação com o custo e uso detalhados. Os grafos por hora mostram os gastos sob demanda com o compromisso recomendado do Savings Plans, fornecendo informações sobre economias estimadas, cobertura do Savings Plans e utilização do Savings Plans. Isso ajuda as organizações a entender como os Savings Plans se aplicam a cada hora de gasto sem precisar investir tempo e recursos na criação de modelos para analisar as despesas.

Crie periodicamente relatórios que contêm um destaque de Savings Plans, instâncias reservadas e recomendações de dimensionamento para o Amazon EC2 do AWS Cost Explorer a fim de 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 ou peça ajuda à equipe de contas e parceiro para 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 para a 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 tendências negativas, bem como tendências positivas que possam ser promovidas em toda a organização. Os relatórios devem envolver representantes das equipes e proprietários de aplicações, bem como do setor financeiro, e os principais tomadores de decisão sobre as despesas com a nuvem.

Recursos

Documentos relacionados:

- AWS Cost Explorer
- AWS Trusted Advisor
- AWS Budgets
- AWS Cost and Usage Report

- Práticas recomendadas do AWS Budgets
- Análises do Amazon S3

Exemplos relacionados:

- Laboratórios do Well-Architected: Governança de custo e uso
- 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: Médio

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: Médio

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

COST01-BP08 Criar uma cultura com reconhecimento de custos

Implemente mudanças ou programas em toda a organização para criar uma cultura com reconhecimento de custos. É recomendável começar aos poucos e, à medida que seus recursos aumentarem e o uso da nuvem por sua organização aumentar, implementar programas grandes e abrangentes.

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

Orientação de implementação

Uma cultura com reconhecimento de custos permite escalar a otimização de custos e o gerenciamento financeiro na nuvem (operações financeiras, centro de excelência da nuvem, equipes de operações na nuvem e assim por diante) por meio de práticas recomendadas executadas de

maneira orgânica e descentralizada em toda a organização. O reconhecimento de custos permite que você crie altos níveis de capacidade em toda a organização com o mínimo de esforço, em comparação com uma abordagem centralizada e de cima para baixo.

Provocar o reconhecimento de custos em computação em nuvem, principalmente para geradores de custos primários na computação em nuvem, permite que as equipes entendam os resultados esperados de quaisquer alterações na perspectiva de custo. As equipes que acessam os ambientes de nuvem devem conhecer os modelos de preços e a diferença entre datacenters on-premises tradicionais e computação em nuvem.

O principal benefício de uma cultura com reconhecimento de custos é que as equipes de tecnologia otimizam os custos de maneira proativa e contínua (por exemplo, são consideradas um requisito não funcional ao arquitetar novas workloads ou alterar workloads existentes) em vez de realizarem otimizações de custo reativas somente quando necessárias.

Pequenas mudanças na cultura podem ter grandes impactos na eficiência de suas cargas de trabalho atuais e futuras. Exemplos disso incluem:

- Oferecer visibilidade e conscientizar as equipes de engenharia para que entendam o que fazem e qual seu impacto em termos de custo.
- Gamificação do custo e do uso em toda a organização. Isso pode ser feito por meio de um painel visível publicamente ou de um relatório que compara custos e uso normalizados entre equipes (por exemplo, custo por workload e custo por transação).
- Reconhecimento da eficiência de custos. Recompense realizações de otimização de custos voluntárias ou não solicitadas publicamente ou de forma privada e aprenda com os erros para evitar repeti-los no futuro.
- Criar requisitos organizacionais de cima para baixo para workloads a serem executadas em orçamentos predefinidos.
- Questionar os requisitos comerciais das mudanças e o impacto sobre os custos das mudanças solicitadas na infraestrutura de arquitetura ou configuração de workload para garantir que você pague somente o necessário.
- Garantir que o planejador de mudanças esteja ciente das mudanças esperadas que impactam o custo, e que sejam confirmadas pelas partes interessadas para que proporcionem resultados comerciais com economia.

Etapas da implementação

- Informar os custos de nuvem às equipes de tecnologia: para conscientizar sobre os custos e estabelecer KPIs de eficiência para partes interessadas financeiras e comerciais.
- Informar partes interessadas ou membros da equipe sobre mudanças planejadas: Crie um item na agenda para discutir mudanças planejadas e o impacto de custo-benefício sobre a workload durante as reuniões semanais de mudanças.
- Reunir-se com a equipe da sua conta: Estabelecer reuniões regulares com a equipe de contas e discutir as tendências do setor e os serviços da AWS. Fale com o gerente de contas, o arquiteto e a equipe de suporte.
- Compartilhe histórias de sucesso: compartilhe histórias de sucesso sobre redução de custo de qualquer workload, Conta da AWS ou organização para gerar uma atitude positiva e encorajar sobre a otimização dos custos.
- Treinamento: garanta que as equipes técnicas ou os membros da equipe sejam treinados reconhecer os custos dos recursos na Nuvem AWS.
- Eventos e reuniões da AWS: participe das conferências locais da AWS e de qualquer reunião local com outras organizações da área.
- Inscrever-se em blogs: acesse as páginas do Blog da AWS e inscreva-se no Blog de novidades e outros blogs relevantes para ficar por dentro dos novos lançamentos, implementações, exemplos e mudanças compartilhados pela AWS.

Recursos

Documentos relacionados:

- Blog da AWS
- Gerenciamento de custos da AWS
- Blog de novidades da AWS

Exemplos relacionados:

- Gerenciamento financeiro na nuvem da AWS
- AWS Well-Architected Labs: gerenciamento financeiro na nuvem

COST01-BP09 Quantificar o valor comercial proveniente da otimização de custos

A quantificação do valor empresarial da otimização de custos permite que você entenda todo o conjunto de benefícios da sua organização. Como a otimização de custos é um investimento necessário, quantificar o valor empresarial permite que você explique o retorno sobre o investimento para as partes interessadas. A quantificação do valor empresarial pode ajudá-lo a ganhar mais participação das partes interessadas em futuros investimentos de otimização de custos e fornece uma estrutura para medir os resultados das atividades de otimização de custos da sua organização.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

Quantificar o valor comercial significa avaliar os benefícios que as empresas obtêm com as ações e decisões que tomam. O valor comercial pode ser tangível (como a redução das despesas ou o aumento dos lucros) ou intangível (como a melhoria da reputação da marca ou o aumento da satisfação do cliente).

Quantificar o valor comercial proveniente da otimização de custos significa determinar o valor ou o benefício que você está obtendo de seus esforços para gastar com maior eficiência. Por exemplo, se uma empresa gastar USD 100 mil para implantar uma workload na AWS e depois otimizá-la, o novo custo se tornará apenas USD 80 mil, sem prejudicar a qualidade ou a produção. Nesse cenário, o valor comercial quantificado da otimização de custos seria uma economia de USD 20 mil. Mas, além das economias, a empresa também pode quantificar o valor em termos de prazos de entrega mais rápidos, maior satisfação do cliente ou outras métricas resultantes das iniciativas de otimização de custos. As partes interessadas precisam tomar decisões sobre o valor em potencial da otimização de custos, o custo da otimização da workload e o valor de retorno.

Além de relatar economias com base na otimização de custos, é recomendável quantificar o valor adicional entregue. Os benefícios de otimização de custos normalmente são quantificados em termos de custos mais baixos por resultado comercial. Por exemplo, é possível quantificar a redução de custo do Amazon Elastic Compute Cloud (Amazon EC2) ao comprar Savings Plans, que reduzem os custos e mantêm os níveis de saída da workload. É possível quantificar a redução de custos em relação aos gastos na AWS quando instâncias ociosas do Amazon EC2 são removidas ou quando volumes não anexados do Amazon Elastic Block Store (Amazon EBS) são excluídos.

No entanto, os benefícios da otimização de custos vão além da redução ou da prevenção de custos. Considere a captura de dados adicionais para medir melhorias de eficiência e valor empresarial.

Etapas da implementação

- Avaliar os benefícios dos negócios: trata-se do processo de analisar e ajustar os custos da Nuvem AWS de forma a maximizar o benefício recebido de cada dólar gasto. Em vez de enfatizar a redução de custos sem valor comercial, considere os benefícios empresarias e o retorno sobre o investimento da otimização de custos, o que pode agregar maior valor ao dispêndio. Isso significa gastar com sabedoria e fazer investimentos e despesas em áreas que geram o melhor retorno.
- Analisar os custos de previsão da AWS: a previsão ajuda as partes interessadas de finanças a definir expectativas com outras partes interessadas internas e externas da organização, além de ajudar a melhorar a previsibilidade financeira da organização. O AWS Cost Explorer pode ser usado para calcular o custo e o uso.

Recursos

Documentos relacionados:

- Nuvem AWS Economics
- Blog da AWS
- Gerenciamento de Custos da AWS
- Blog de novidades da AWS
- whitepaper sobre o Pilar Confiabilidade do Well-Architected
- Explorador de Custos da AWS

Vídeos relacionados:

- Desbloquear o valor comercial com o Windows na AWS

Exemplos relacionados:

- Medir e maximizar o valor comercial do Customer 360
- The Business Value of Adopting Amazon Web Services Managed Databases
- The Business Value of Amazon Web Services for Independent Software Vendors
- Business Value of Cloud Modernization
- The Business Value of Migration to Amazon Web Services

Reconhecimento de despesas e usos

Perguntas

- CUSTOS 2. Como governar o uso?
- CUSTOS 3. Como monitorar custos e uso?
- CUSTOS 4. Como desativar os recursos?

CUSTOS 2. Como governar 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 Desenvolver políticas com base nos requisitos da sua organização
- COST02-BP02 Implementar objetivos e metas
- COST02-BP03 Implementar uma estrutura de contas
- COST02-BP04 Implementar grupos e perfis
- COST02-BP05 Implementar controles de custos
- COST02-BP06 Acompanhar o ciclo de vida do projeto

COST02-BP01 Desenvolver políticas com base nos requisitos da sua organização

Desenvolva políticas que definam como os recursos são gerenciados pela sua organização e os inspecione periodicamente. As políticas devem abranger aspectos de custos de recursos e workloads, 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 para implementação

Entender os custos e os motivadores da sua organização é essencial para gerenciar seus custos e utilização com eficácia e também 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 de negócio organização, cada uma com o próprio fluxo de

receita. A capacidade de atribuir os custos dos recursos a cargas de trabalho, à uma organização em particular ou aos donos do produto propicia um comportamento eficiente de uso e ajuda a reduzir o desperdício. O monitoramento preciso de custos e uso ajuda você a entender a otimização de sua workload, bem como a lucratividade das unidades e produtos da organização. Esse conhecimento permite uma tomada de decisão mais consciente sobre onde alocar recursos em 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 manter ciente do seu custo e utilização.

O primeiro passo para realizar governança é usar os requisitos da sua organização para desenvolver políticas para o uso da nuvem. Essas políticas definem como a sua organização usa a nuvem e como os recursos são gerenciados. As políticas devem abranger todos os aspectos de recursos e workloads relacionados ao custo ou à utilização, incluindo criação, modificação e desativação durante a vida útil do recurso. Verifique se as políticas e procedimentos são seguidos e implementados para qualquer alteração em um ambiente de nuvem. Durante as reuniões de gestão de mudanças de TI, questione para descobrir o impacto do custo das alterações planejadas, sejam de aumento ou diminuição, a justificativa de negócios e o resultado esperado.

As políticas devem ser simples, para que sejam facilmente compreendidas e possam ser implementadas com eficácia em toda a organização. As políticas também precisam ser fáceis de seguir e interpretar (para que sejam usadas) e específicas (para evitar erros de interpretação entre as equipes). Além disso, elas precisam ser inspecionadas periodicamente (como nossos mecanismos) e atualizadas à medida que as condições ou as prioridades de negócios dos clientes mudam, o que tornaria a política desatualizada.

Comece com políticas amplas e de alto nível, como qual região geográfica usar 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 dados com menor performance em ambientes de teste e desenvolvimento), quais tipos de recursos podem ser usados por diferentes grupos (por exemplo, o maior tamanho de um recurso em uma conta de desenvolvimento é médio) e por quanto tempo esses recursos ficarão em uso (se temporariamente, em curto prazo ou por um período específico).

Exemplo de política

Veja a seguir um exemplo de política que você pode revisar para criar suas próprias políticas de governança de nuvem, que enfocam a otimização de custos. Ajuste a política com base nos requisitos de sua organização e nas solicitações das partes interessadas.

- Nome da política: Defina um nome de política claro, como Política de otimização de recursos e redução de custos.
- Finalidade: Explique por que essa política deve ser usada e qual é o resultado esperado. O objetivo dessa política é verificar se há um custo mínimo necessário para implantar e executar a workload desejada para atender aos requisitos de negócios.
- Escopo: Defina claramente quem deve usar essa política e quando ela deve ser usada, como o DevOps X Team, para usar essa política em clientes do leste dos EUA para o ambiente X (produção ou não produção).

Declaração de política

1. Selecione us-east-1 ou várias regiões do leste dos EUA com base no ambiente de sua workload e nos requisitos de negócios (desenvolvimento, teste de aceitação do usuário, pré-produção ou produção).
2. Programe instâncias do Amazon EC2 e do Amazon RDS para execução entre 6h e 20h (Horário Padrão do Leste (EST)).
3. Interrompa todas as instâncias do Amazon EC2 não utilizadas após oito horas e as instâncias do Amazon RDS não utilizadas após 24 horas de inatividade.
4. Encerre todas as instâncias do Amazon EC2 não utilizadas após 24 horas de inatividade em ambientes que não sejam de produção. Lembre o proprietário da instância do Amazon EC2 (com base em tags) de revisar suas instâncias do Amazon EC2 interrompidas em produção e informá-lo de que suas instâncias do Amazon EC2 serão encerradas em 72 horas se não estiverem em uso.
5. Use família de instância e tamanho genéricos, como m5.large, e, depois, redimensione a instância com base na utilização da CPU e da memória usando o AWS Compute Optimizer.
6. Priorize o uso do ajuste de escala automático para ajustar dinamicamente o número de instâncias em execução com base no tráfego.
7. Use instâncias spot para workloads não essenciais.
8. Analise os requisitos de capacidade para comprometer Saving Plans ou instâncias reservadas para workloads previsíveis e informe a equipe de gerenciamento financeiro da nuvem.
9. Use políticas de ciclo de vida do Amazon S3 para mover dados acessados com pouca frequência para níveis de armazenamento mais baratos. Se nenhuma política de retenção for definida, use o Amazon S3 Intelligent Tiering para mover objetos automaticamente para a camada arquivada.
10. Monitore a utilização de recursos e defina alarmes para acionar eventos de escalabilidade usando o Amazon CloudWatch.

11. Para cada Conta da AWS, use o AWS Budgets para definir orçamentos de custo e uso para sua conta com base no centro de custos e nas unidades de negócios.
12. Usar o AWS Budgets para definir orçamentos de custo e uso para sua conta pode ajudar você a controlar seus gastos e evitar contas inesperadas, permitindo controlar melhor seus custos.

Procedimento: Forneça procedimentos detalhados para implementar essa política ou consulte outros documentos que descrevam como implementar cada declaração de política. Esta seção deve fornecer instruções detalhadas para a elaboração dos requisitos da política.

Para implementar essa política, você pode usar várias ferramentas de terceiros ou regras do AWS Config para conferir a conformidade com a declaração de política e acionar ações de correção automatizadas usando funções do AWS Lambda. Você também pode usar o AWS Organizations para aplicar a política. Além disso, você deve revisar regularmente o uso de recursos e ajustar a política conforme necessário para verificar se ela continua atendendo às suas necessidades comerciais.

Etapas da implementação

- Reúna-se com as partes interessadas: Para desenvolver políticas, peça às partes interessadas (escritórios de negócios na nuvem, engenheiros ou tomadores de decisão funcionais para aplicação de políticas) em sua organização que especifiquem seus requisitos e os documentem. Adote uma abordagem ampla e iterativa iniciando em alto nível com um refinamento contínuo até os mínimos detalhes em cada etapa. Os membros da equipe incluem aqueles com interesse direto na workload, como unidades da organização ou proprietários de aplicativos, bem como grupos de apoio, como equipes de segurança e finanças.
- Obtenha confirmação: garanta que as equipes concordem com as políticas de quem pode acessar e implantar na Nuvem AWS. Certifique-se de que elas sigam as políticas da sua organização e confirme se o provisionamento de recursos está alinhado com as políticas e procedimentos estabelecidos.
- Crie sessões de treinamento de integração: peça que os novos membros completem cursos de formação de integração para criar conscientização de custo e requisitos da organização. As equipes podem assumir diferentes políticas das suas experiências anteriores ou nem pensar nelas.
- 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 as 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 como 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 workload para garantir que permaneça alinhada com as prioridades da organização.
- 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:

- Gerenciamento de alterações na nuvem
- Políticas gerenciadas pela AWS para funções de trabalho
- Estratégia de faturamento de várias contas da AWS
- Ações, recursos e chaves de condição para serviços da AWS
- Gerenciamento e governança da AWS
- Controlar o acesso às Regiões da AWS usando as políticas do IAM
- Regiões e AZs de infraestruturas globais

Vídeos relacionados:

- AWS Management and Governance at Scale (Gerenciamento e governança da AWS em grande escala)

Exemplos relacionados:

- VMware: o que são políticas de nuvem?

COST02-BP02 Implementar objetivos e metas

Implemente objetivos e metas de custo e uso para sua workload. Os objetivos fornecem orientação para sua organização quanto aos resultados esperados, e as metas oferecem resultados mensuráveis específicos a serem alcançados para suas workloads.

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

Orientação para implementação

Desenvolva objetivos e metas de custo e uso para a sua organização. Como uma organização em crescimento na AWS, é importante definir e monitorar objetivos para otimização de custos. Esses objetivos ou indicadores-chave de desempenho (KPIs) podem incluir itens como porcentagem de gastos sob demanda ou adoção de certos serviços otimizados, como instâncias do AWS Graviton ou tipos de volume gp3 do EBS. Definir objetivos mensuráveis e viáveis pode ajudar você a continuar a medir as melhorias de eficiência, o que é importante para as operações comerciais em andamento. Os objetivos fornecem orientações e direcionamento para a sua organização sobre os resultados esperados. As metas fornecem resultados mensuráveis específicos a serem alcançados. Em suma, um objetivo é a direção que você deseja seguir e a meta é até que ponto nessa direção o objetivo deve ir e quando ele deve ser concretizado (usando a orientação de específico, mensurável, atribuível, realista e oportuno, ou SMART). Um exemplo de objetivo é que 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 workloads precisam ser mais eficientes a cada seis meses. A meta complementar seria o custo de acordo com as métricas empresariais diminuir em 5% a cada seis meses.

Uma meta para a otimização de custos é aumentar a eficiência da workload, o que significa diminuir o custo por resultado empresarial da workload ao longo do tempo. É recomendável implementar esse objetivo para todas as workloads e também definir uma meta, como um aumento de 5% na eficiência

a cada seis meses a um ano. Isso pode ser obtido na nuvem por meio da criação de recursos na otimização de custos e do lançamento de serviços e recursos.

É importante ter visibilidade quase em tempo real sobre seus KPIs e oportunidades de economia relacionadas e acompanhar seu progresso ao longo do tempo. Para começar a definir e monitorar os objetivos de KPI, recomendamos o painel de KPI da framework de Painéis de inteligência em nuvem (CID). Com base nos dados do AWS Cost and Usage Report, o painel de KPI oferece uma série de KPIs de otimização de custos recomendados com a capacidade de definir objetivos personalizados e acompanhar o progresso ao longo do tempo.

Se você tiver outra solução que possibilite definir e monitorar objetivos de KPI, garanta que ela seja adotada por todos as partes interessadas de gerenciamento financeiro em nuvem em sua organização.

Etapas da implementação

- Definir níveis de uso esperados: Para começar, enfoque os níveis de uso. Interaja com os proprietários de aplicativos, 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 haverá alterações devido a aumentos sazonais ou campanhas de marketing?
- Definir custos e recursos de workload: Com os níveis de uso definidos, quantifique as alterações nos recursos da workload necessárias para atender a esses níveis de uso. Pode ser necessário aumentar o tamanho ou o número de recursos para um componente de workload, aumentar a transferência de dados ou alterar componentes de workload 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.
- Definir objetivos empresariais: Combine o resultado das alterações esperadas no uso e no custo com as alterações esperadas na tecnologia ou qualquer programa que você esteja executando e desenvolva metas para a carga de trabalho. Os objetivos devem abordar o uso, o custo e a relação entre os dois. Os objetivos devem ser simples, de alto nível e ajudar as pessoas a entenderem o que o negócio espera em termos de resultados (como garantir que recursos não utilizados sejam mantidos abaixo de determinado nível de custo). Não é necessário definir objetivos para cada tipo de recurso não utilizado ou definir custos que causem perdas para objetivos e metas. Verifique se há programas organizacionais (por exemplo, criação de recursos como treinamento e educação) se houver alterações esperadas no custo sem alterações no uso.
- Definir metas: Para cada uma das metas definidas, especifique um objetivo mensurável. Se o objetivo for aumentar a eficiência na workload, a meta quantificará a melhoria (típica nos

resultados de negócios para cada dólar gasto) e quando ela será entregue. Por exemplo, se você define um objetivo de minimizar o desperdício causado pelo superprovisionamento, sua meta pode ser que o desperdício decorrente do superprovisionamento de computação no primeiro nível de workloads de produção não exceda 10% do custo de computação do nível e que o desperdício devido ao superprovisionamento de computação no segundo nível de workloads de produção não exceda 5% do custo de computação do nível.

Recursos

Documentos relacionados:

- Políticas gerenciadas pela AWS para funções de trabalho
- Estratégia de várias contas da AWS para sua zona de pouso do AWS Control Tower
- Controlar o acesso às Regiões da AWS usando as políticas do IAM
- Objetivos do SMART

Vídeos relacionados:

- Laboratórios do Well-Architected: Objetivos e metas (nível 100)

Exemplos relacionados:

- Laboratórios do Well-Architected: Recursos de desativação (objetivos e metas)
- Laboratórios do Well-Architected: Tipo, tamanho e número do recurso (objetivos e metas)

COST02-BP03 Implementar uma estrutura de contas

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 exposição a riscos quando esta prática recomendada não é estabelecida: alto

Orientações para a implementação

O AWS Organizations permite criar várias Contas da AWS que podem ajudar você a gerenciar de maneira centralizada seu ambiente à medida que dimensiona suas workloads na AWS. É possível modelar sua hierarquia organizacional agrupando Contas da AWS na estrutura da unidade

organizacional (UO) e criando várias Contas da AWS em cada UO. Para criar uma estrutura de contas, primeiramente, você precisa decidir qual das suas Contas da AWS será a conta de gerenciamento. Depois disso, você pode criar Contas da AWS ou selecionar contas existentes como contas membro com base na estrutura de contas projetada seguindo as práticas recomendadas de conta de gerenciamento e as práticas recomendadas de conta membro.

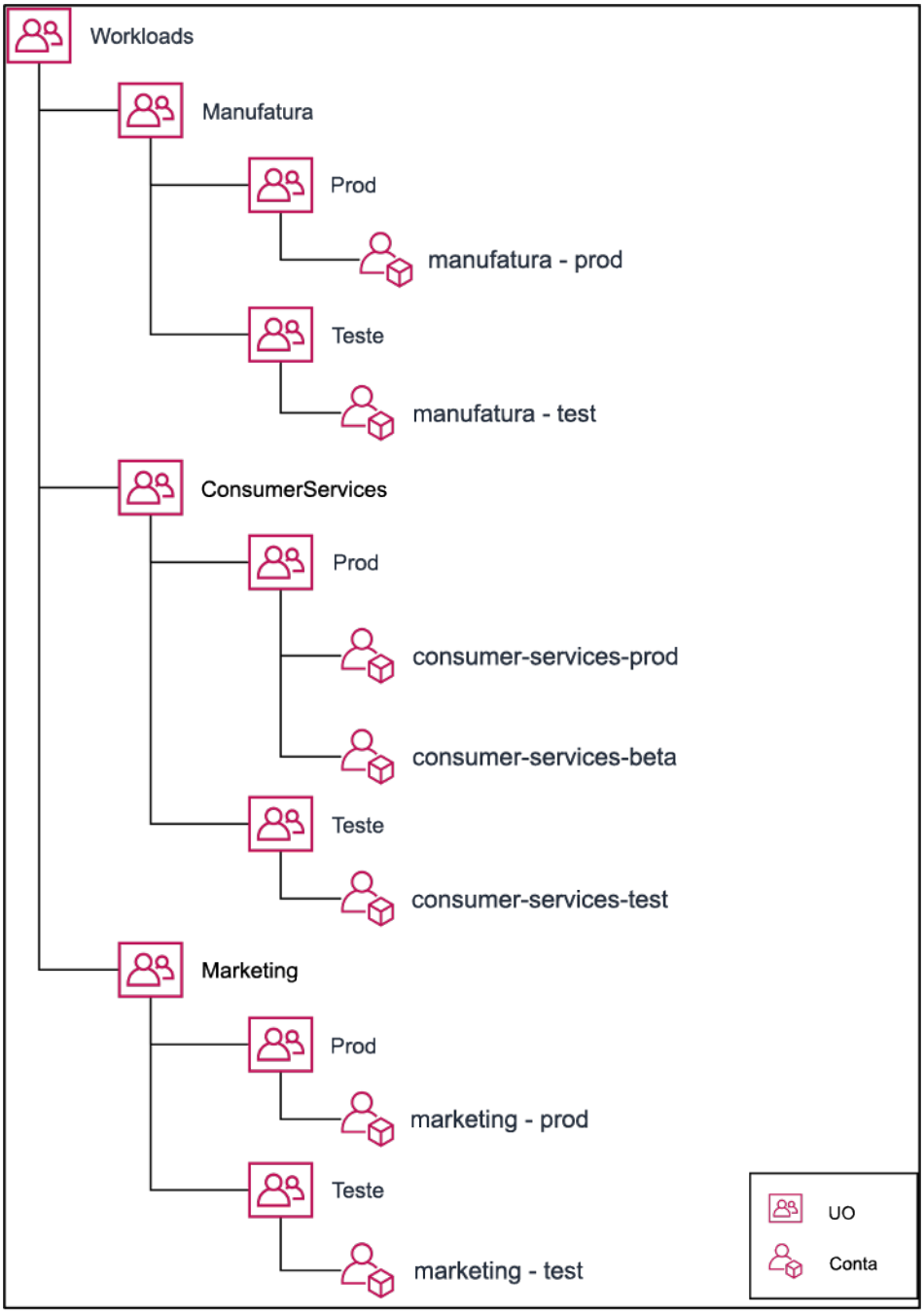
Recomenda-se que sempre haja, pelo menos, uma conta de gerenciamento com uma conta membro, independentemente do tamanho ou uso da organização. Todos os recursos de workload devem residir somente nas contas membro e nenhum recurso deve ser criado na conta de gerenciamento. Não há uma resposta geral para a quantidade de Contas da AWS que 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 de negócios, por exemplo:

- O isolamento administrativo ou fiscal e de faturamento é necessário entre unidades da organização, centros de custo ou workloads específicas.
- Os limites de serviço da AWS são definidos para que sejam específicos a determinadas workloads.
- Há um requisito de isolamento e separação entre workloads e recursos.

Dentro do AWS Organizations, o faturamento consolidado cria a estrutura entre uma ou mais contas membro 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 workload (workload 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 da AWS membro 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, você pode 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.

O diagrama a seguir mostra como você pode usar o AWS Organizations com unidades organizacionais (UO) para agrupar várias contas e colocar várias Contas da AWS em cada UO. Recomenda-se usar UOs para vários casos de uso e workloads que fornecem padrões para organizar contas.



Exemplo de agrupamento de várias Contas da AWS em unidades organizacionais.

AWS Control Tower pode instalar e configurar rapidamente várias contas da AWS, garantindo que a governança esteja alinhada com os requisitos da organização.

Etapas da implementação

- Definir requisitos de separação: os requisitos de separação são uma combinação de vários fatores, que incluem estruturas de segurança, de confiabilidade e financeiras. Trabalhe em cada fator em

ordem e especifique se a workload ou o ambiente dela deve ser separado de outras workloads. A segurança promove a adesão aos requisitos de acesso e de dados. A confiabilidade gerencia os limites para que os ambientes e as workloads não afetem os outros. Revise os pilares de segurança e de confiabilidade do Well-Architected Framework periodicamente e siga as práticas recomendadas fornecidas. As estruturas financeiras criam separação financeira rígida (diferentes centros de custo, propriedades de workload e responsabilidades). Exemplos comuns de separação são workloads de produção e de teste executadas em contas separadas ou o uso de uma conta separada para que os dados da fatura e do faturamento possam ser fornecidos às unidades de negócios individuais ou aos departamentos da organização, ou à parte interessada que possui a conta.

- Definir requisitos de agrupamento: os requisitos de agrupamento não substituem os requisitos de separação, mas são usados para auxiliar no gerenciamento. Agrupe ambientes semelhantes ou workloads que não exigem separação. Um exemplo disso é o agrupamento de vários ambientes de teste ou desenvolvimento de uma ou mais workloads.
- Definir a estrutura de contas: Usando essas separações e agrupamentos, especifique uma conta para cada grupo e mantenha os requisitos de separação. Essas contas são suas contas membro ou vinculadas. Ao agrupar essas contas membro 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 membro uma visualização individual dos dados de faturamento. Se uma conta membro não precisar ter os dados de uso ou de faturamento visíveis para nenhuma outra conta, ou se uma fatura separada da AWS for necessária, você deverá definir 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 membro ou vinculadas. As contas de gerenciamento ou pagantes devem ser usadas somente para gerenciamento.

Recursos

Documentos relacionados:

- Uso de tags de alocação de custos
- Políticas gerenciadas da AWS para funções de trabalho
- Estratégia de faturamento de várias contas da AWS
- Controle o acesso a Regiões da AWS usando as políticas do IAM
- AWS Control Tower

- AWS Organizations
- Práticas recomendadas para contas de gerenciamento e contas membro
- Organização do ambiente usando várias contas da AWS
- Ativação de instâncias reservadas compartilhadas e descontos de Savings Plans
- Faturamento consolidado
- Faturamento consolidado

Exemplos relacionados:

- Divisão do CUR e compartilhamento do acesso

Vídeos relacionados:

- Apresentação do AWS Organizations
- Configuração de um ambiente de várias contas da AWS que usa as práticas recomendadas para AWS Organizations

Exemplos relacionados:

- Laboratórios do Well-Architected: criação de uma organização da AWS (Nível 100)
- Divisão do AWS Cost and Usage Report e compartilhamento do acesso
- Definição de uma estratégia de várias contas da AWS para empresas de telecomunicações
- Práticas recomendadas para otimização das Contas da AWS
- Práticas recomendadas para unidades organizacionais com o AWS Organizations

COST02-BP04 Implementar grupos e perfis

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 a serviços da AWS e a soluções de terceiros.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: baixo

Orientações para a implementação

Os perfis e os grupos de usuários são elementos fundamentais no design e na implementação de sistemas seguros e eficientes. Os perfis e os grupos ajudam as organizações a equilibrar a necessidade de controle com a necessidade de flexibilidade e produtividade, e, acima de tudo, apoiando os objetivos organizacionais e as necessidades dos usuários. Conforme recomendado na seção Gerenciamento de identidade e acesso do Pilar Segurança: AWS Well-Architected Framework, você precisa ter permissões e gerenciamento de identidade robustos para fornecer acesso aos recursos certos para as pessoas certas nas condições certas. Os usuários recebem somente o acesso necessário para realizar suas tarefas. Isso minimiza o risco associado ao acesso não autorizado ou ao uso indevido.

Depois de desenvolver políticas, é possível criar perfis e grupos lógicos de usuários em sua organização. Isso permite que você atribua permissões, controle o uso e ajude a implementar mecanismos robustos de controle de acesso, impedindo o acesso não autorizado a informações sigilosas. Comece com agrupamentos de pessoas de alto nível. Normalmente isso se alinha com as unidades organizacionais e os cargos (por exemplo, administrador de sistemas no departamento de TI, controlador financeiro ou analista de negócios). Os grupos categorizam pessoas que realizam tarefas semelhantes e precisam de acesso semelhante. As funções definem o que um grupo deve fazer. É mais fácil gerenciar permissões para grupos e perfis do que para usuários individuais. Os perfis e os grupos atribuem permissões de forma consistente e sistemática a todos os usuários, evitando erros e inconsistências.

Quando o perfil de um usuário muda, os administradores podem ajustar o acesso por perfil ou grupo, em vez de reconfigurar as contas de usuários individuais. 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

- Implementar grupos: usando os grupos de usuários definidos em suas políticas organizacionais, implemente os grupos correspondentes, se necessário. Para obter as práticas recomendadas sobre usuários, grupos e autenticação, consulte o Pilar Segurança: AWS Well-Architected Framework.
- Implementar perfis e políticas: usando as ações definidas nas políticas organizacionais, crie os perfis e as políticas de acesso necessárias. Para obter as práticas recomendadas sobre perfis e políticas, consulte o Pilar Segurança: AWS Well-Architected Framework.

Recursos

Documentos relacionados:

- Políticas gerenciadas da AWS para funções de trabalho
- Estratégia de faturamento de várias contas da AWS
- Pilar Segurança: AWS Well-Architected Framework
- AWS Identity and Access Management (IAM)
- Políticas do AWS Identity and Access Management

Vídeos relacionados:

- Why use Identity and Access Management

Exemplos relacionados:

- Laboratório do Well-Architected: Identidade e acesso básico
- Control access to Regiões da AWS using IAM policies
- Starting your Cloud Financial Management journey: Cloud cost operations

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, como controle do acesso a regiões ou tipos de recursos.

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

Orientações para a implementação

Uma primeira etapa comum na implementação de controles de custo é configurar notificações quando eventos de custo ou de uso ocorrerem fora das políticas. É possível tomar medidas rápidas e verificar se é necessária uma ação corretiva, sem restringir ou afetar negativamente workloads ou novas atividades. Depois de conhecer os limites de workload e do ambiente, você pode aplicar a governança. O AWS Budgets permite que você defina notificações e orçamentos mensais para custos, uso e descontos de compromisso da conta 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.

Depois de configurar seus limites de orçamento com o AWS Budgets, use AWS Cost Anomaly Detection para reduzir seu custo inesperado. AWS Cost Anomaly Detection é um serviço de gerenciamento de custos que usa machine learning para monitorar continuamente o custo e o uso a fim de detectar gastos incomuns. Ele ajuda a identificar gastos anômalos e causas raiz, para que você possa agir rapidamente. Primeiro, crie um monitor de custos em AWS Cost Anomaly Detection e depois escolha sua preferência de alerta configurando um limite em dólares (como um alerta sobre anomalias com impacto superior a USD 1 mil). Com o recebimento dos alertas, é possível analisar a causa raiz por trás da anomalia e o impacto em seus custos. Também é possível monitorar e realizar sua própria análise de anomalias em AWS Cost Explorer.

Aplique políticas de governança na AWS por meio de AWS Identity and Access Management e AWS Organizations Políticas de controle de serviço (SCPs). O IAM permite gerenciar com segurança o acesso a serviços e recursos da AWS. Usando o IAM, você pode controlar quem pode criar ou gerenciar recursos da AWS, os tipos de recursos que podem ser criados e onde eles podem ser criados. Isso minimiza a possibilidade de criação de recursos fora da política definida. 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. Para ter mais detalhes sobre a implementação do gerenciamento de acesso, consulte o whitepaper Pilar Segurança do Well-Architected.

A governança também pode ser implementada por meio do gerenciamento do Service Quotas da AWS. Ao garantir que o Service Quotas esteja 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 suas políticas de organização definidas, crie AWS Budgets para ser notificado 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.

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. Você também pode pré-configurar ações de orçamento do AWS que podem aplicar políticas específicas de IAM ou SCP ou interromper Amazon EC2 de destino ou instâncias de Amazon RDS. As ações de orçamento podem ser executadas automaticamente ou exigir aprovação do fluxo de trabalho.

- Implementar notificações sobre gastos: Use o AWS Cost Anomaly Detection para reduzir seus custos imprevistos em sua organização e analisar a causa raiz de possíveis gastos anômalos. Depois de criar o monitor de custos para identificar gastos incomuns em sua granularidade especificada e configurar notificações em AWS Cost Anomaly Detection, ele envia um alerta quando um gasto incomum é detectado. Isso permitirá que você analise o caso raiz por trás da anomalia e entenda o impacto em seu custo. Use categorias de custo de AWS durante a configuração de AWS Cost Anomaly Detection para identificar qual equipe de projeto ou equipe de unidade de negócios pode analisar a causa raiz do custo inesperado e tomar as ações necessárias em tempo hábil.
- Implementar controles de uso: Usando as políticas da organização definidas, implemente políticas e perfis do IAM para especificar quais ações os usuários podem e quais 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 as políticas do IAM
- AWS Budgets
- AWS Cost Anomaly Detection
- Controle de custos do AWS

Vídeos relacionados:

- Como posso usar o AWS Budgets para rastrear meus gastos e uso

Exemplos relacionados:

- Exemplos de políticas de gerenciamento de acesso do IAM
- Políticas de controle de serviço de exemplo
- Ações de orçamento do AWS
- Criar política do IAM para controlar o acesso aos recursos do Amazon EC2 usando tags
- Restringir o acesso do Identity do IAM a recursos específicos do Amazon EC2
- Criar uma política do IAM para restringir o uso do Amazon EC2 por família
- Laboratórios do Well-Architected: Governança de custo e uso (Nível 100)
- Laboratórios do Well-Architected: Governança de custo e uso (Nível 200)
- Integrações do Slack para Cost Anomaly Detection usando AWS Chatbot

COST02-BP06 Acompanhar o ciclo de vida do projeto

Acompanhe, meça e realize auditorias no ciclo de vida dos projetos, equipes e ambientes para evitar o uso e pagamento de recursos desnecessários.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: baixo

Orientações para a implementação

O monitoramento eficaz do ciclo de vida do projeto permite que as organizações tenham um controle mais adequado sobre os custos por meio de melhor planejamento, gerenciamento e otimização de recursos, tempo e qualidade. Os insights obtidos por meio do rastreamento são inestimáveis para a tomada de decisões fundamentadas que contribuem para a relação custo-benefício e o sucesso geral do projeto.

O rastreamento de todo o ciclo de vida da workload ajuda a compreender quando as workloads ou os respectivos componentes não são mais necessários. As workloads e os componentes existentes podem parecer estar em uso, mas quando a AWS libera novos serviços ou recursos, eles podem ser desativados ou adotados. Confira os estágios anteriores das workloads. Depois que uma workload está em produção, os ambientes anteriores podem ser desativados ou terem a capacidade significativamente reduzida 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. É possível usar o AWS Config ou o AWS Systems Manager para fornecer um inventário detalhado da configuração e dos seus recursos 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 de eventos e métricas avançados fornecido pela AWS permite criar uma visão de eventos de ciclo de vida significativos e gerenciar os recursos proativamente para reduzir os custos desnecessários.

De modo semelhante ao gerenciamento do ciclo de vida da aplicação (ALM), o acompanhamento do ciclo de vida do projeto deve envolver vários processos, ferramentas e equipes que trabalham juntas, como design e desenvolvimento, testes, produção, suporte e redundância de workload.

Ao monitorar cuidadosamente cada fase do ciclo de vida de um projeto, as organizações obtêm insights cruciais e um controle aprimorado, o que facilita o sucesso do planejamento, da implementação e da conclusão do projeto. Essa supervisão cuidadosa verifica se os projetos, além de atenderem aos padrões de qualidade, são entregues no prazo e dentro do orçamento, promovendo o custo-benefício de modo geral.

Para obter mais informações sobre como implementar o rastreamento do ciclo de vida de entidades, consulte o whitepaper Pilar Excelência operacional: AWS Well-Architected.

Etapas da implementação

- Estabelecer um processo de monitoramento do ciclo de vida de projetos: a equipe do Centro de Excelência da Nuvem deve estabelecer o processo de monitoramento do ciclo de vida de projetos. Estabeleça uma abordagem estruturada e sistemática para monitorar as workloads a fim de melhorar o controle, a visibilidade e o desempenho dos projetos. Torne o processo de monitoramento transparente, colaborativo e dedicado à melhoria contínua para maximizar sua eficácia e valor.
- Realizar análises da workload: conforme definido por suas políticas organizacionais, configure uma frequência regular para auditar os projetos existentes e realizar análises da workload. 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:

- Guidance for Tagging on AWS
- O que é ALM (gerenciamento do ciclo de vida das aplicações)?
- Políticas gerenciadas da AWS para funções de trabalho

Exemplos relacionados:

- Control access to Regiões da AWS using IAM policies

Ferramentas relacionadas:

- AWS Config
- AWS Systems Manager
- AWS Budgets
- AWS Organizations
- AWS CloudFormation

CUSTOS 3. Como monitorar custos e uso?

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

Práticas recomendadas

- COST03-BP01 Configurar fontes de informações detalhadas
- COST03-BP02 Adicionar informações da organização ao custo e ao uso
- COST03-BP03 Identificar categorias de atribuição de custos
- COST03-BP04 Estabelecer métricas da organização
- COST03-BP05 Configurar as ferramentas de faturamento e gerenciamento de custos
- COST03-BP06 Alocar custos com base nas métricas de workload

COST03-BP01 Configurar fontes de informações detalhadas

Configure as ferramentas de gerenciamento de custos e geração de relatórios para detalhamento por hora a fim de fornecer informações detalhadas de custo e uso, bem como aumentar o nível de análise e transparência. Configure a workload para gerar ou receber as entradas de log para cada resultado empresarial entregue.

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

Orientação para implementação

Informações detalhadas de faturamento, como detalhamento por hora nas ferramentas de gerenciamento de custos, permitem que as organizações acompanhem suas taxas de consumo com mais detalhes e as ajudem a identificar alguns dos motivos do aumento de custos. Essas fontes de dados oferecem a visualização mais precisa do custo e do uso em toda a organização.

O AWS Cost and Usage Report fornece detalhamento 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
- Detalhamento por hora
- Versionamento: Substituir relatório existente
- Integração de dados: Athena (formato Parquet e compactação)

Use 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 níveis de detalhamento para fornecer informações resumidas de alto nível.

- Configurar a granularidade por hora no Cost Explorer: Habilite o Por hora e Dados em nível de recurso para acessar dados de custo e uso com granularidade por hora nos últimos 14 dias e granularidade em nível de recurso.
- Configurar o registro em log de aplicações: Verifique se a aplicação registra cada resultado empresarial entregue para que possa ser acompanhado e medido. Verifique se o detalhamento desses dados é pelo menos por hora para que ele corresponda aos dados de custo e uso. Para obter mais detalhes sobre registro em log e monitoramento, consulte Pilar Excelência operacional do Well-Architected.

Recursos

Documentos relacionados:

- AWS Cost and Usage Report
- 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 de AWS Cost and Usage Reports
- Pilar Excelência operacional do Well-Architected

Exemplos relacionados:

- Configuração da conta da AWS
- O novo visual e os casos de uso comuns do AWS Cost Explorer

COST03-BP02 Adicionar informações da organização ao custo e ao uso

Defina um esquema de marcação com base na sua organização, atributos da workload e categorias de alocação de custos para que você possa filtrar e pesquisar recursos ou monitorar custos e uso em ferramentas de gerenciamento de custos. Implemente marcação consistente em todos os recursos, sempre que possível, por finalidade, equipe, ambiente ou outros critérios relevantes ao seu negócio.

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

Orientações para a implementação

Implemente 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 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 de organização (como centros de custo, nomes de aplicativos, 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 aos 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 as políticas de tags no AWS Organizations para definir regras de como as tags podem ser usadas em recursos da AWS nas suas contas no AWS Organizations. As políticas de tag permitem que você adote facilmente uma abordagem padronizada para marcar os recursos da AWS.

O Tag Editor da AWS permite adicionar, excluir e gerenciar tags de vários recursos. Com o Tag Editor, é possível pesquisar os recursos que você deseja marcar e gerenciar as tags para os recursos nos resultados da pesquisa.

As Categorias de Custos da AWS permitem que você atribua significado da organização 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

- Defina um esquema de marcação: reúna todas as partes interessadas de todo o seu negócio 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.

- Recursos de tag: usando suas categorias de atribuição de custos definidas, coloque tags em todos os recursos em suas workloads de acordo com as categorias. Use ferramentas como CLI, Tag Editor ou AWS Systems Manager para aumentar a eficiência.
- Implemente as Categorias de Custos da AWS: você pode criar categorias de custos sem implementar a marcação. As categorias de custos usam as dimensões de custo e uso existentes. Crie regras de categoria a partir do esquema e as implemente nas categorias de custos.
- Automatize 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 serviços como o AWS CloudFormation para garantir que os recursos sejam marcados quando forem criados. Você também pode criar uma solução personalizada para marcar automaticamente usando as funções do Lambda ou usar um microsserviço que verifica a workload periodicamente e remove todos os recursos que não estão marcados, o que é ideal para ambientes de teste e desenvolvimento.
- Monitore e relate a 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:

- Práticas recomendadas de marcação
- Tag de recurso do AWS CloudFormation
- Categorias 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

Vídeos relacionados:

- How can I tag my AWS resources to divide up my bill by cost center or project (Como posso marcar meus recursos da AWS para dividir minha fatura por centro de custo ou projeto)
- Marcação de recursos da AWS

Exemplos relacionados:

- Marcar automaticamente novos recursos da AWS com base em identidade ou perfil

COST03-BP03 Identificar categorias de atribuição de custos

Identifique categorias organizacionais, como unidades de negócios, departamentos ou projetos, que poderiam ser usadas para alocar custos em sua organização às entidades consumidoras internas. Use essas categorias para impor a responsabilidade de gastos, bem como promover o reconhecimento de custos e comportamentos de consumo eficazes.

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

Orientação para implementação

O processo de categorização de custos é crucial em orçamentos, contabilidade, relatórios financeiros, tomada de decisão, benchmarking e gerenciamento de projetos. Ao classificar e categorizar as despesas, as equipes podem entender melhor os tipos de custos que gerarão ao longo da jornada para a nuvem, ajudando-as a tomar decisões conscientes e gerenciar orçamentos de forma eficaz.

A responsabilidade pelos gastos com a nuvem estabelece um forte incentivo para o gerenciamento disciplinado da demanda e dos custos. O resultado é uma economia significativamente maior nos custos da nuvem para organizações que alocam a maior parte de seus gastos com a nuvem para unidades de negócios ou equipes consumidoras. Além disso, a alocação de gastos na nuvem ajuda as organizações a adotar mais práticas recomendadas de governança centralizada da nuvem.

Trabalhe com sua equipe financeira e outras partes interessadas relevantes para entender os requisitos de como os custos devem ser alocados em sua organização durante suas chamadas regulares. Os custos da workload 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.

Depois de definir as categorias de atribuição de custos com as partes interessadas na organização, use Categorias de custos da AWS para agrupar as informações de custo e uso em categorias significativas na Nuvem AWS, como custo de um projeto específico ou em Contas da AWS para departamentos ou unidades de negócios. É possível criar categorias personalizadas e mapear as informações de custo e uso nessas categorias com base nas regras definidas usando várias dimensões, como conta, tag, serviço ou tipo de cobrança. Assim que as categorias de custos forem definidas, você verá as informações de custos e uso de acordo com elas, permitindo que a organização tome melhores decisões estratégicas e de compras. Também é possível ver essas categorias no AWS Cost Explorer, no AWS Budgets e no AWS Cost and Usage Report.

Por exemplo, é possível criar categorias de custos para suas unidades de negócios (equipe DevOps) e, em cada categoria, criar várias regras (para cada subcategoria) com várias dimensões (Contas da AWS, tags de alocação de custos, serviços ou tipo de cobrança) com base nos seus agrupamentos definidos. Com as categorias de custos, é possível organizar os custos usando um mecanismo baseado em regras. As regras que você configurar organizarão seus custos em categorias. Dentro dessas regras, é possível aplicar filtros usando várias dimensões para cada categoria, como Contas da AWS, serviços da AWS ou tipos de cobrança específicos. Depois, você pode usar essas categorias em vários produtos no console do AWS Billing and Cost Management e Console de Gerenciamento de Custos .. Isso inclui AWS Cost Explorer, AWS Budgets, AWS Cost and Usage Report e AWS Cost Anomaly Detection.

Como exemplo, o diagrama a seguir mostra como agrupar as informações de custos e uso em sua organização, com várias equipes (categoria de custos), vários ambientes (regras) e vários recursos ou ativos em cada ambiente (dimensões).

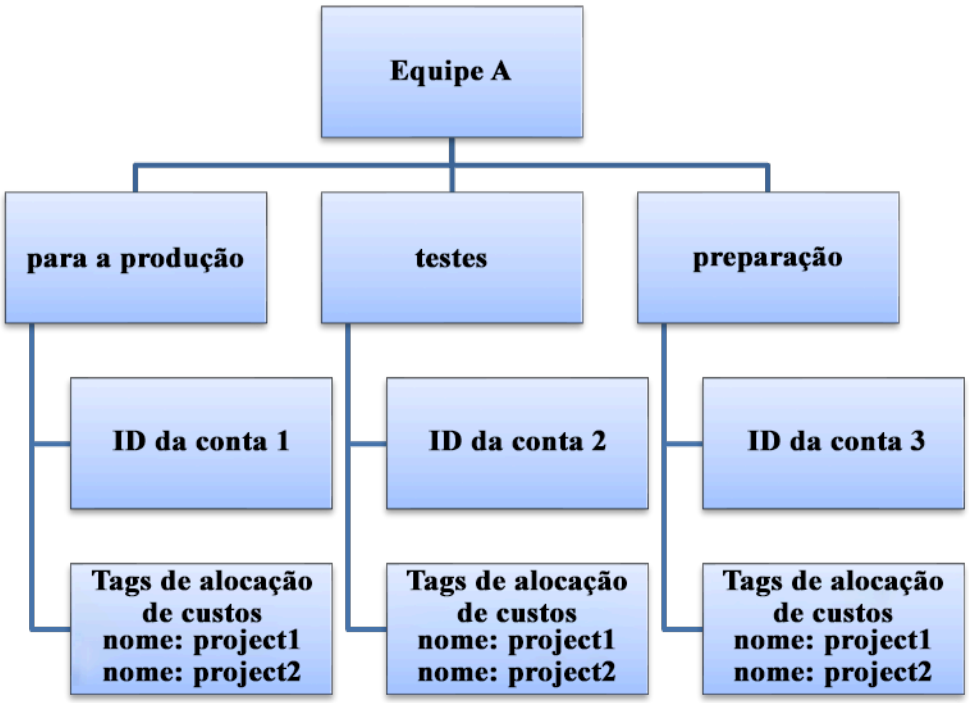


Gráfico de organização de custos e uso

Também é possível criar agrupamento de custos usando as categorias de custos. Depois de criar as categorias de custos (aguardando até 24 horas após a criação de uma categoria para que seus registros de uso sejam atualizados com valores), elas aparecem no AWS Cost Explorer, o AWS Budgets, o AWS Cost and Usage Reporte o AWS Cost Anomaly Detection. No AWS Cost Explorer e no AWS Budgets, uma categoria de custos aparece como uma dimensão de faturamento adicional. Você pode usar isso para filtrar por valor de categoria de custos específico ou agrupar pela categoria de custos.

Etapas para a implementação

- Defina as categorias da sua organização: Reúna-se com as unidades de negócios e as partes interessadas internas para definir categorias que reflitam a estrutura e os requisitos da organização. Essas categorias devem ser associadas diretamente à 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.
- Defina suas categorias funcionais: Reúna-se com as unidades de negócios e as partes interessadas internas para definir categorias que reflitam as funções presentes na empresa.

Podem ser os nomes da workload ou do aplicativo e o tipo de ambiente, como produção, teste ou desenvolvimento.

- Defina as Categorias de custos da AWS: Crie categorias de custo para organizar as informações de custo e uso usando Categorias de custos da AWS e associe o custo e o uso de recursos da AWS a categorias significativas. Várias categorias podem ser atribuídas a um recurso, e um recurso pode estar em várias categorias diferentes. Portanto, defina quantas categorias forem necessárias para gerenciar seus custos dentro da estrutura categorizada usando Categorias de custos da AWS.

Recursos

Documentos relacionados:

- Marcação de recursos da AWS
- Uso de tags de alocação de custos
- Análise de custos com o AWS Budgets
- Análise de custos com o Cost Explorer
- Gerenciamento de AWS Cost and Usage Reports
- Categorias de custos da AWS
- Gerenciamento de custos com as Categorias de custos da AWS
- Criação de categorias de custos
- Marcação de categorias de custos
- Divisão de cobranças entre as categorias de custos
- Recursos das Categorias de custos da AWS

Exemplos relacionados:

- Organizar seus dados de custos e uso com as Categorias de custos da AWS
- Gerenciamento de custos com as Categorias de custos da AWS
- Laboratórios do Well-Architected: Visualização de custo e uso
- Laboratórios do Well-Architected: Categorias de custo

COST03-BP04 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 é 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

- Defina os resultados da workload: reúna-se com as partes interessadas do negócio e defina os resultados para a workload. 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.
- Defina os resultados para os componentes da workload: opcionalmente, se você tiver uma workload grande e complexa ou puder facilmente dividir sua workload 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-BP05 Configurar as ferramentas de faturamento e gerenciamento de custos

Configure as ferramentas de gerenciamento de custos de acordo com as políticas da sua organização para gerenciar e otimizar gastos com a nuvem. Isso inclui serviços, ferramentas e recursos para organizar e rastrear dados de custos e uso, aprimorar o controle por meio de faturamento consolidado e permissão de acesso, melhorar o planejamento por meio de orçamento e previsões, receber notificações ou alertas e reduzir ainda mais os custos com recursos e otimizações de preços.

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

Orientação para implementação

Para estabelecer uma forte responsabilidade, sua estratégia de conta deve ser considerada primeiro como parte de sua estratégia de alocação de custos. Realize essa tarefa do jeito certo e talvez não precise ir além. Caso contrário, poderá haver inconsciência e outros pontos problemáticos.

Para incentivar a responsabilidade pelos gastos com a nuvem, os usuários devem ter acesso a ferramentas que forneçam visibilidade de seus custos e uso. É recomendável que todas as workloads e equipes tenham as ferramentas configuradas para os seguintes detalhes e finalidades:

- Organização: Estabeleça a alocação de custos e linha de base de governança com sua própria estratégia de marcação e categorizações.
- Organização: estabeleça a alocação de custos e linha de base de governança com sua própria estratégia de marcação e taxonomia. Marque os recursos compatíveis da AWS e categorize-os de uma maneira que faça sentido com base na estrutura da sua organização (unidades de negócios, departamentos ou projetos). Marque nomes de contas para centros de custo específicos e mapeie-os com Categorias de Custos da AWS para agrupar contas de unidades de negócios específicas nos centros de custo, para que o proprietário da unidade de negócios possa ver o consumo de várias contas em um só lugar.
- Acesso: acompanhe as informações de faturamento de toda a organização no faturamento consolidado e verifique se as partes interessadas e os proprietários de negócios certos têm acesso.
- Controle: Crie mecanismos de governança eficazes com as barreiras de proteção certas para evitar cenários inesperados ao usar políticas de controle de serviços (SCPs), políticas de tags e alertas de orçamento. Por exemplo, você pode permitir que as equipes criem recursos nas regiões preferidas usando somente mecanismos de controle eficazes.

- 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. Você pode usar o Painel de inteligência em nuvem (CID) ou qualquer outro produto compatível para criar essa visibilidade.
- Notificações: forneça notificações quando o custo ou o uso estiverem fora dos limites definidos e quando ocorrerem anomalias com o AWS Budgets ou o AWS Cost Anomaly Detection.
- Relatórios: Resuma todas as informações de custos e uso e aumente a conscientização e a responsabilidade sobre seus gastos com a nuvem com dados de custos atribuíveis, detalhados e alocáveis. Os relatórios devem ser relevantes para a equipe que os consome e, preferencialmente, devem conter recomendações.
- Rastreamento: mostra o custo e o uso atuais em relação a metas ou objetivos configurados.
- Análises: permita que os membros da equipe realizem análises personalizadas e detalhadas até a granularidade horária, com todas as dimensões possíveis.
- Inspeção: mantenha-se atualizado com suas oportunidades de implantação de recursos e otimização de custos. Receba notificações (usando o Amazon CloudWatch, o Amazon SNS ou o Amazon SES) para implantações de recursos na organização e analise as recomendações de otimização de custos (por exemplo, AWS Compute Optimizer ou AWS Trusted Advisor).
- Tendências: exiba a variabilidade de custo e uso ao longo do período necessário, com a granularidade necessária.
- Previsões: mostre os custos futuros estimados, faça uma estimativa do uso de recursos e gaste com painéis de previsão criados por você.

Você pode usar ferramentas da AWS, como AWS Cost Explorer, o AWS Billing and Cost Managementou AWS Budgets para o essencial, ou você pode integrar dados CUR com Amazon Athena e o Amazon QuickSight para fornecer esse recurso para visualizações mais detalhadas. Se você não tem habilidades essenciais ou largura de banda em sua organização, pode trabalhar com o AWS ProServ, AWS Managed Services (AMS)ou AWS Partners e usar suas ferramentas. Você também pode usar ferramentas de terceiros. Porém, verifique primeiro se o custo agrega valor à sua organização.

Etapas para a implementação

- Permita o acesso baseado em equipe às ferramentas: Configure suas contas e crie grupos que tenham acesso aos relatórios de custo e uso necessários para o consumo e o uso AWS Identity and Access Management para controlar o acesso a ferramentas, como o AWS Cost Explorer.

Esses grupos devem incluir representantes de todas as equipes que possuem ou gerenciam um aplicativo. Isso garante que cada equipe tenha acesso às próprias informações de custo e uso para rastrear seu consumo.

- Configurar o AWS Budgets: Configure o AWS Budgets em todas as contas das workloads. Defina orçamentos para o gasto geral da conta e orçamentos para as workloads usando tags. Configure notificações no AWS Budgets para receber alertas quando você exceder valores orçados ou quando seus custos estimados excederem seus orçamentos.
- Configure o AWS Cost Explorer: Configure o AWS Cost Explorer para sua workload e contas para visualizar seus dados de custos para análise posterior. Crie um painel para a workload que rastreie o gasto geral, as principais métricas de uso da workload e a previsão de custos futuros com base nos seus dados de custo históricos.
- Configure o AWS Cost Anomaly Detection: Use o AWS Cost Anomaly Detection para as contas, os serviços centrais ou as categorias de custos criadas para monitorar os custos e o uso e detectar gastos incomuns. Você pode receber alertas individualmente em relatórios agregados, assim como alertas por e-mail ou em um tópico do Amazon SNS, o que permite analisar e determinar a causa-raiz de uma anomalia e identificar o fator que está aumentando o custo.
- Configure ferramentas avançadas: Como opção, você pode criar ferramentas personalizadas para a organização que forneçam detalhes e granularidade adicionais. Você pode implementar o recurso de análise avançada usando o Amazon Athenae painéis usando o Amazon QuickSight. Pense no uso da Solução CID que tem painéis pré-configurados e avançados. Há também AWS Partners com quem você pode trabalhar e adotar as soluções de gerenciamento de nuvem deles para habilitar o monitoramento e a otimização de faturas de nuvem em um local conveniente.

Recursos

Documentos relacionados:

- Gerenciamento de custos da AWS
- Marcação Recursos da AWS
- Análise de custos com o AWS Budgets
- Análise de custos com o Cost Explorer
- Managing AWS Cost and Usage Report
- Categorias de custos da AWS
- Gerenciamento financeiro na nuvem com a AWS

- Políticas de controle de serviço de exemplo
- AWS APN Partners – Cost Management

Vídeos relacionados:

- Deploying Cloud Intelligence Dashboards (Implantação de painéis de inteligência de nuvem)
- Get Alerts on any FinOps or Cost Optimization Metric or KPI (Receber alertas sobre qualquer FinOps ou métrica de otimização de custos ou KPI)

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
- Laboratórios do Well-Architected: Painéis de inteligência de nuvem
- Como usar SCPs para definir barreiras de proteção de permissão nas contas

COST03-BP06 Alocar custos com base nas métricas de workload

Aloque os custos da workload por métricas de uso ou resultados de negócios para medir a eficiência de custos da workload. Implemente um processo para analisar os dados de custo e uso com serviços de análise, que podem fornecer informações e capacidade de estorno.

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

Orientação para implementação

A otimização de custos está fornecendo resultados de negócios com o menor preço, que só pode ser alcançado ao alocar custos de workload por métricas de workload (medidas pela eficiência da workload). 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 de maneira permanente) 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.

Criar métricas de workload para recursos compartilhados pode ser um desafio em comparação a recursos, como aplicações em contêineres no Amazon Elastic Container Service (Amazon ECS) e no Amazon API Gateway. No entanto, existem algumas maneiras de categorizar o uso e rastrear os custos. Se precisar monitorar recursos compartilhados do Amazon ECS e do AWS Batch, você poderá habilitar os dados de alocação de custos divididos no AWS Cost Explorer. Com dados de alocação de custos divididos, você pode entender e otimizar o custo e o uso de suas aplicações em contêineres e alocar os custos das aplicações para entidades comerciais individuais com base na forma como os recursos compartilhados de computação e memória são consumidos. Se você compartilhou o uso da função do API Gateway e do AWS Lambda, poderá usar o AWS Application Cost Profiler para categorizar seu consumo com base em seu `ID do locatário` ou `ID do cliente`.

Etapas da implementação

- Aloque custos para métricas de workload: Usando as métricas definidas e a tags configuradas, crie uma métrica que combine a saída e o custo da workload. Use serviços de análise, 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

Exemplos relacionados:

- Melhore a visibilidade de custos do Amazon ECS e do AWS Batch com dados de alocação de custos divididos da AWS.

CUSTOS 4. Como desativar os recursos?

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

Práticas recomendadas

- COST04-BP01 Acompanhar os recursos ao longo da vida útil
- COST04-BP02 Implementar um processo de desativação
- COST04-BP03 Desativar recursos
- COST04-BP04 Desativar recursos automaticamente
- COST04-BP05 Reforçar políticas de retenção de dados

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 workload ou a função do recurso.

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

Orientações para a implementação

Desative recursos de workload que não são mais necessários. Um exemplo comum são os recursos usados para testes: após a conclusão do teste, os recursos podem ser removidos. O rastreamento de recursos com tags (e execução de relatórios sobre essas tags) pode ajudar a identificar ativos para desativação, pois eles não estarão em uso ou a licença deles expirará. 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 feature-X` para identificar a finalidade do recurso em termos de ciclo de vida da workload. Outro exemplo consiste em usar o `LifeSpan` ou `TTL` para os recursos, como nome e valor da chave de tag a ser excluída para definir o período ou o tempo específico para desativação.

Etapas da implementação

- Implementar 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. A marcação ajuda a categorizar os recursos por finalidade,

equipe, ambiente ou outros critérios relevantes para o seu negócio. Para obter mais detalhes sobre casos de uso, estratégias e técnicas de marcação, consulte Práticas recomendadas de marcação de AWS.

- Implementar o monitoramento da saída ou do throughput da workload: Implemente monitoramento ou alarme de throughput da workload, acionando solicitações de entrada ou conclusões de saída. Configure-o para fornecer notificações quando saídas ou solicitações de workload caírem para zero, indicando que os recursos de workload não são mais usados. Incorpore um fator de tempo se a workload cair periodicamente para zero em condições normais. Para obter mais detalhes sobre recursos não utilizados ou subutilizados, consulte Verificações de otimização de custos de AWS Trusted Advisor.
- Agrupar recursos do AWS: Crie grupos para recursos do AWS. Você pode usar o AWS Resource Groups para organizar e gerenciar seus recursos do AWS que estão no mesmo Região da AWS. Você pode adicionar tags à maioria de seus recursos para ajudar a identificá-los e classificá-los em sua organização. Use Tag Editor para adicionar tags aos recursos compatíveis em massa. Considere o uso de AWS Service Catalog para criar, gerenciar e distribuir portfólios de produtos aprovados para usuários finais e gerenciar o ciclo de vida do produto.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor
- Verificações de otimização de custos do AWS Trusted Advisor
- Marcação de recursos da AWS
- Publicar métricas personalizadas

Vídeos relacionados:

- Como otimizar custos usando AWS Trusted Advisor

Exemplos relacionados:

- Organização de recursos de AWS
- Otimização do custo usando AWS Trusted Advisor

COST04-BP02 Implementar um processo de desativação

Implemente um processo para identificar e desativar recursos não utilizados.

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

Orientações para a 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 verificar se todos os requisitos da organização foram atendidos.

Etapas da implementação

- Criar e implementar um processo de desativação: Trabalhe com os proprietários e desenvolvedores de workloads e crie um processo de desativação para a workload e os recursos dela. O processo deve abranger o método para verificar se a workload está em uso e também se cada um dos recursos da workload está em uso. Detalhe 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, devem ser incluídos. Notifique os proprietários da workload de que o processo de desativação foi executado.

  Use as seguintes etapas de desativação para obter orientações sobre o que deve ser verificado como parte do seu processo:

  - Identificar os recursos a serem desativados: Identifique os recursos elegíveis para desativação em seu Nuvem AWS. Registre todas as informações necessárias e agende a desativação. Em sua linha do tempo, certifique-se de considerar se (e quando) problemas inesperados surgirem durante o processo.
  - Coordenar e comunicar: Trabalhe com os proprietários da workload para confirmar o recurso a ser desativado
  - Registrar metadados e criar backups: Registre metadados (como IPs públicos, região, AZ, VPC, sub-rede e grupos de segurança) e crie backups (como snapshots do Amazon Elastic Block Store ou obtenção de AMI, exportação de chaves e exportação de certificado) se for necessário para os recursos no ambiente de produção ou se forem recursos críticos.
  - Validar a infraestrutura como código: Determine se os recursos foram implantados com o AWS CloudFormation, Terraform, AWS Cloud Development Kit (AWS CDK) ou qualquer outra ferramenta de implantação de infraestrutura como código para que possam ser reimplantados, se necessário.

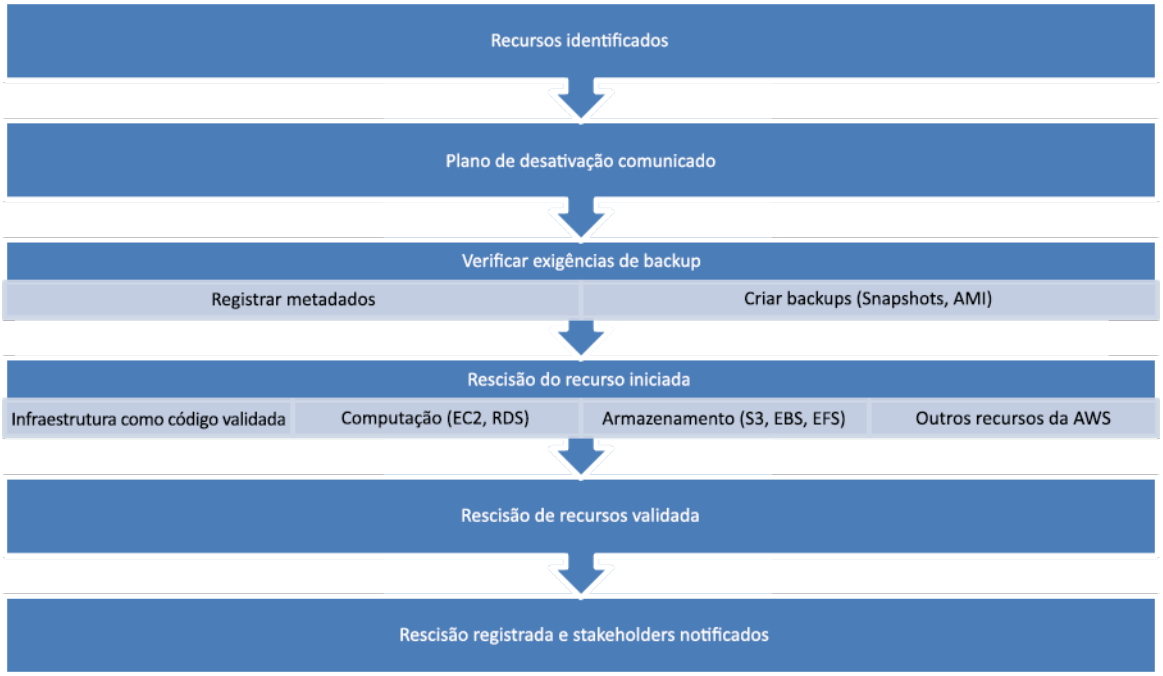
- Impedir acesso: Aplique controles restritivos por um período, para evitar o uso de recursos enquanto você determina se o recurso é necessário. Verifique se o ambiente de recursos pode ser revertido para seu estado original, se necessário.
- Seguir seu processo de desativação interno: Siga as tarefas administrativas e o processo de desativação de sua organização, como remover o recurso do domínio da organização, remover o registro DNS e remover o recurso de sua ferramenta de gerenciamento de configuração, ferramenta de monitoramento, ferramenta de automação e ferramentas de segurança.

Se o recurso for uma instância do Amazon EC2, consulte a lista a seguir. Para obter mais detalhes, consulte Como faço para excluir ou terminar meus recursos do Amazon EC2?

- Interrompa ou encerre todas as suas instâncias do Amazon EC2 e balanceadores de carga. As instâncias do Amazon EC2 ficam visíveis no console por um curto período após serem finalizadas. Você não será cobrado por instâncias que não estiverem em estado de execução
- Exclua sua infraestrutura do Auto Scaling.
- Libere todos os hosts dedicados.
- Exclua todos os volumes do Amazon EBS e snapshots do Amazon EBS.
- Libere todos os endereços IP elásticos.
- Cancele o registro das imagens de máquina da Amazon (AMIs).
- Encerre todos os ambientes do AWS Elastic Beanstalk.

Se o recurso for um objeto armazenado no Amazon S3 Glacier e se você excluir um arquivo antes de atingir a duração mínima de armazenamento, será cobrada uma taxa proporcional de exclusão antecipada. A duração mínima do armazenamento do Amazon S3 Glacier depende da classe de armazenamento usada. Para obter um resumo da duração mínima de armazenamento para cada classe de armazenamento, consulte Desempenho nas classes de armazenamento do Amazon S3. Para obter detalhes sobre como as taxas de exclusão antecipada são calculadas, consulte Definição de preço do Amazon S3.

O fluxograma simples do processo de desativação a seguir descreve as etapas de desativação. Antes de desativar recursos, verifique se os recursos que você identificou para desativação não estão sendo usados pela organização.



Fluxo de desativação de recursos.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor
- AWS CloudTrail

Vídeos relacionados:

- Excluir pilha do CloudFormation, mas reter alguns recursos
- Descubra qual usuário iniciou a instância do Amazon EC2

Exemplos relacionados:

- Excluir ou terminar recursos do Amazon EC2
- Descubra qual usuário iniciou uma instância do Amazon EC2

COST04-BP03 Desativar recursos

Desative recursos acionados por eventos, como auditorias periódicas ou alterações no uso. Em geral, a desativação pode ser realizada periodicamente e é manual ou automatizada.

Nível de risco exposto se esta prática recomendada não é 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 workload, 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: Esta é a fase de depreciação de recursos do AWS que não são mais necessários ou término de um contrato de licenciamento. Conclua todas as verificações finais concluídas antes de passar para o estágio de descarte e desativação de recursos para evitar interrupções indesejadas, como tirar snapshots ou backups. Usando o processo de desativação, desative cada um dos recursos que foram identificados como não utilizados.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor

Exemplos relacionados:

- Laboratórios do Well-Architected: Recursos de desativação (Nível 100)

COST04-BP04 Desativar recursos automaticamente

Projete a workload 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 exposição a riscos quando esta prática recomendada não é estabelecida: baixo

Orientações para a implementação

Use a automação para reduzir ou remover os custos associados do processo de desativação. Projetar sua workload para executar a desativação automatizada reduzirá os custos gerais da workload durante sua vida útil. É possível usar o AWS Auto Scaling para realizar o processo de desativação. Você também pode implementar código personalizado usando a API ou o SDK para desativar recursos de workload automaticamente.

As aplicações modernas são desenvolvidas sem servidor como prioridade, uma estratégia que prioriza a adoção de serviços sem servidor. O AWS desenvolveu serviços sem servidor para todas as três camadas de sua pilha: computação, integração e armazenamento de dados. O uso da arquitetura sem servidor permitirá que você economize custos durante períodos de baixo tráfego com aumento e redução automáticos.

Etapas da implementação

- Implementar o AWS Auto Scaling: Para recursos compatíveis, configure-os com o AWS Auto Scaling. O AWS Auto Scaling pode ajudar você a otimizar sua utilização e eficiência de custos ao consumir serviços do AWS. Quando a demanda cair, o AWS Auto Scaling removerá automaticamente qualquer excesso de capacidade de recursos para evitar gastos excessivos.
- Configurar o CloudWatch para encerrar instâncias: As instâncias podem ser configuradas para encerrar usando alarmes de CloudWatch. Usando as métricas do processo de desativação, implemente um alarme com uma ação do Amazon Elastic Compute Cloud. Verifique a operação em um ambiente que não seja de produção antes de implantar.
- Implementar o código na workload: Você pode usar o SDK do AWS ou AWS CLI para desativar recursos de workload. Implemente código dentro da aplicação que se integra à AWS e encerre ou remova recursos não mais usados.
- Usar serviços sem servidor: Priorize a criação de arquiteturas sem servidor e arquitetura baseada em eventos no AWS para criar e executar seus aplicativos. O AWS oferece vários serviços de tecnologia sem servidor que fornecem inerentemente a utilização de recursos otimizada automaticamente e a desativação automatizada (aumentar e reduzir a escala horizontalmente). Com aplicativos sem servidor, a utilização de recursos é otimizada automaticamente e você nunca paga por provisionamento em excesso.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- AWS Trusted Advisor
- Tecnologia sem servidor no AWS
- Crie alarmes para interromper, encerrar, reinicializar ou recuperar uma instância
- Conceitos básicos do Amazon EC2 Auto Scaling
- Adição de ações de encerramento para alarmes do Amazon CloudWatch

Exemplos relacionados:

- Agendamento de exclusão automática de pilhas do AWS CloudFormation
- Laboratórios do Well-Architected: Recursos de desativação automática (Nível 100)
- Limpeza automatizada da AWS/Servian

COST04-BP05 Reforçar políticas de retenção de dados

Defina as políticas de retenção de dados em recursos compatíveis para lidar com exclusão de objetos de acordo com os requisitos de suas organizações. Identifique e exclua recursos e objetos desnecessários ou órfãos que não sejam mais necessários.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Use políticas de retenção de dados e de ciclo de vida para reduzir os custos associados do processo de desativação e de armazenamento dos recursos identificados. A definição de suas políticas de retenção de dados e de ciclo de vida para realizar a exclusão e a migração automatizadas de classe de armazenamento reduzirá os custos gerais de armazenamento durante seu tempo de vida. Você pode usar o Amazon Data Lifecycle Manager para automatizar a criação e a exclusão de snapshots do Amazon Elastic Block Store e imagens de máquina (AMIs) baseadas no Amazon EBS, e o Amazon S3 Intelligent-Tiering ou uma configuração de ciclo de vida do Amazon S3 para gerenciar o ciclo de vida de seus objetos do Amazon S3. Você também pode implementar código personalizado usando a API ou o SDK para criar políticas de ciclo de vida e regras de políticas para objetos a serem excluídos de forma automática.

Etapas da implementação

- Uso do Amazon Data Lifecycle Manager: use políticas de ciclo de vida no Amazon Data Lifecycle Manager para automatizar a exclusão de snapshots do Amazon EBS e AMIs baseadas no Amazon EBS.
- Definição da configuração do ciclo de vida em um bucket: use a configuração do ciclo de vida do Amazon S3 em um bucket para definir ações a serem realizadas pelo Amazon S3 durante o ciclo de vida de um objeto, bem como a exclusão no final do ciclo de vida do objeto, com base nos requisitos de sua empresa.

Recursos

Documentos relacionados:

- AWS Trusted Advisor
- Amazon Data Lifecycle Manager
- How to set lifecycle configuration on Amazon S3 bucket (Como definir a configuração de ciclo de vida em um bucket do Amazon S3)

Vídeos relacionados:

- Automate Amazon EBS Snapshots with Amazon Data Lifecycle Manager (Automatizar snapshots do Amazon EBS com o Amazon Data Lifecycle Manager)
- Empty an Amazon S3 bucket using a lifecycle configuration rule (Esvaziar um bucket do Amazon S3 com o uso de uma regra de configuração de ciclo de vida)

Exemplos relacionados:

- Empty an Amazon S3 bucket using a lifecycle configuration rule (Esvaziar um bucket do Amazon S3 com o uso de uma regra de configuração de ciclo de vida)
- Laboratório do Well-Architected: Recursos de desativação automática (Nível 100)

Recursos econômicos

Perguntas

- CUSTOS 5. Como avaliar o custo ao selecionar serviços?
- CUSTOS 6. Como atingir as metas de custo ao selecionar tamanho, número e tipo de recurso?

- CUSTOS 7. Como usar os modelos de definição de preço para reduzir custos?
- CUSTOS 8. Como planejar as cobranças de transferência de dados?

CUSTOS 5. Como avaliar 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 os requisitos de custos da organização
- COST05-BP02 Analisar todos os componentes da workload
- COST05-BP03 Executar uma análise completa de cada componente
- COST05-BP04 Selecionar software com licenciamento econômico
- COST05-BP05 Selecionar os componentes desta workload para otimizar o custo alinhado com as prioridades da organização
- COST05-BP06 Realizar análises de custos para diferentes usos ao longo do tempo

COST05-BP01 Identificar os requisitos de custos da organização

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 exposição a riscos se esta prática recomendada não for estabelecida: alto

Orientações para a implementação

Na maioria das organizações, o departamento de tecnologia da informação (TI) é composto de várias equipes pequenas, cada uma com sua própria agenda e área de foco, que refletem as especialidades e as habilidades dos respectivos membros. Você precisa compreender os objetivos, as prioridades e as metas gerais da organização e como cada departamento ou projeto contribui para esses objetivos. A categorização de todos os recursos essenciais, incluindo pessoal, equipamentos, tecnologia, materiais e serviços externos, é crucial para alcançar os objetivos organizacionais e um planejamento orçamentário abrangente. A adoção dessa abordagem

sistemática para a identificação e a compreensão dos custos é fundamental para estabelecer um plano de custos realista e robusto para a organização.

ao selecionar serviços para a sua carga de trabalho, é fundamental compreender as prioridades da sua organização. Crie um equilíbrio entre a otimização de custos e outros pilares do AWS Well-Architected Framework, como desempenho e confiabilidade. Esse processo deve ser conduzido de forma sistemática e regular para refletir as mudanças nos objetivos da organização, nas condições de mercado e na dinâmica operacional. 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 organização, como produtos, negócios, técnicas e finanças, para coletar as informações. Avalie o impacto das compensações entre interesses concorrentes ou abordagens alternativas para ajudar a tomar decisões fundamentadas ao determinar onde concentrar as iniciativas 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 a aplicação.

Etapas da implementação

- Identificar as necessidades de custos da organização: reúna-se com os membros das equipes da organização, incluindo aqueles em gerenciamento de produtos, proprietários de aplicações, equipes de desenvolvimento e de operações, gerenciamento e finanças. Priorize os pilares do Well-Architected para essa workload e os respectivos componentes. O resultado deve ser uma lista ordenada dos pilares. Também é possível adicionar um peso a cada pilar para indicar quanto foco adicional ele tem ou quão semelhantes são os focos entre dois pilares.
- Abordar a dívida técnica e documentá-la: durante a análise da workload, aborde a dívida técnica. Documente um item de backlog para revisitar a workload no futuro, com o objetivo de refatorar ou rearquitetar para otimizá-la ainda mais. É essencial comunicar claramente as compensações feitas para outras partes interessadas.

Recursos

Práticas recomendadas relacionadas:

- REL11-BP07 Arquitetar o produto para cumprir as metas de disponibilidade e os acordos de nível de serviço (SLAs) de tempo de atividade

- OPS01-BP06 Avalie as compensações

Documentos relacionados:

- AWS Total Cost of Ownership (TCO) Calculator (Calculadora de custo total de propriedade (TCO) da AWS)
- Categorias de armazenamento do Amazon S3
- Produtos da nuvem

COST05-BP02 Analisar todos os componentes da workload

Verifique se cada componente da workload é analisado, independentemente do tamanho ou dos custos atuais. O trabalho da análise deve refletir o benefício em potencial, como os custos atuais e projetados.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: alto

Orientações para a implementação

Os componentes da workload, projetados para agregar valor comercial à organização, podem abranger vários serviços. Para cada componente, é possível escolher serviços específicos da Nuvem AWS para atender às necessidades dos negócios. Essa seleção pode ser influenciada por fatores como a familiaridade ou a experiência anterior com esses serviços.

Depois de identificar os requisitos da organização (conforme mencionado em COST05-BP01 Identificar os requisitos de custos da organização), faça uma análise completa de todos os componentes da workload. Analise cada componente considerando os custos e os tamanhos atuais e projetados. Considere o custo da análise em relação a qualquer possível economia da workload ao longo do respectivo ciclo de vida. O trabalho despendido na análise de todos os componentes dessa workload deve corresponder às possíveis economias ou melhorias previstas da otimização desse componente específico. Por exemplo, se o custo do recurso proposto for USD 10/mês e, sob as cargas previstas, não excederem USD 15/mês, gastar um dia de trabalho para reduzir os custos em 50% (USD 5 por mês) poderá exceder o benefício em potencial durante a vida útil do sistema. O uso de 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, e o conjunto certo de serviços poderá 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. Por exemplo, o uso do Amazon EMR Serverless pode ser a escolha apropriada inicialmente. No entanto, à medida que o consumo desse serviço aumenta, a transição para o Amazon EMR no Amazon EC2 pode reduzir os custos desse componente da workload.

A análise estratégica de todos os componentes da workload, independentemente de seus atributos atuais, tem o potencial de gerar melhorias notáveis e economias financeiras ao longo do tempo. O esforço investido nesse processo de análise deve ser deliberado, com consideração cuidadosa das vantagens que podem ser obtidas.

O AWS Cost Explorer e o AWS Cost and Usage Report (CUR) podem analisar o custo de uma prova de conceito (PoC) ou do ambiente em execução. Também é possível usar o AWS Pricing Calculator para estimar os custos da workload.

Etapas da implementação

- Listar os componentes da workload: crie uma lista dos componentes da workload. Ela é usada para a verificação de que cada componente foi analisado. O esforço despendido deve refletir a criticidade da workload conforme definido pelas prioridades da organização. O agrupamento dos recursos de forma funcional melhora a eficiência (por exemplo, o armazenamento dos bancos de dados de produção, se houver vários bancos de dados).
- Priorizar a lista de componentes: priorize a lista de componentes em ordem de esforço. Normalmente, isso é feito por ordem de custos dos componentes, do mais caro para o mais barato, ou da criticidade, conforme definido pelas prioridades da organização.
- Executar a análise: para cada componente da lista, analise as opções e os serviços disponíveis e escolha a opção mais alinhada com as suas prioridades organizacionais.

Recursos

Documentos relacionados:

- AWS Pricing Calculator
- AWS Cost Explorer
- Categorias de armazenamento do Amazon S3
- Produtos da nuvem

COST05-BP03 Executar uma análise completa de cada componente

Observe o custo geral de cada componente para a organização. Calcule o custo total de propriedade considerando o custo de operações e gerenciamento, especialmente ao usar serviços gerenciados pelo provedor de nuvem. 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 é estabelecida: Alto

Orientações para a implementação

Considere a economia de tempo que permitirá que sua equipe se concentre na retirada de recursos de endividamento técnico, inovação, agregação de valor e criação do que diferencia os negócios. Por exemplo, talvez você precise mover sem alterações (lift-and-shift) seu ambiente on-premises para a nuvem (também conhecido como redefinir a hospedagem) e otimizar mais tarde. Vale a pena explorar as possíveis economias obtidas com o uso de serviços gerenciados na AWS que removem ou reduzem os custos de licença. Serviços gerenciados na AWS eliminam a sobrecarga operacional e administrativa da manutenção de um serviço, como aplicação de patches ou atualização do sistema operacional, e permitem que você se concentre na inovação e nos negócios.

Uma vez que os serviços gerenciados operam em escala da nuvem, eles podem oferecer menor custo por transação ou serviço. Você pode realizar possíveis otimizações para alcançar alguns benefícios tangíveis, sem alterar a arquitetura principal da aplicação. Por exemplo, você pode tentar reduzir o tempo gasto no gerenciamento de instâncias de banco de dados migrando para uma plataforma de banco de dados como serviço, como o Amazon Relational Database Service (Amazon RDS), ou migrando sua aplicação para uma plataforma totalmente gerenciada, como o AWS Elastic Beanstalk.

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 ativar 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 workloads 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 workload. 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.

O Amazon RDS, o Amazon Redshift e o Amazon ElastiCache fornecem um serviço de banco de dados gerenciado. O Amazon Athena, o Amazon EMR e o 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

- Realização de 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.
- Comparação de recursos gerenciados e não gerenciados: considere o custo operacional dos recursos que você gerencia e compare-os com recursos gerenciados pela AWS. Por exemplo, analise seus bancos de dados em execução em instâncias do Amazon EC2 e compare-os com as opções do Amazon RDS (um serviço gerenciado pela AWS) ou do Amazon EMR em comparação com a execução do Apache Spark no Amazon EC2. Ao migrar de uma workload autogerenciada para uma workload totalmente gerenciada pela AWS, pesquise suas opções com cuidado. Os três fatores mais importantes a serem considerados são o tipo de serviço gerenciado que você deseja usar, o processo utilizado para migrar seus dados e o entendimento do modelo de responsabilidade compartilhada da AWS.

Recursos

Documentos relacionados:

- AWS Total Cost of Ownership (TCO) Calculator (Calculadora de custo total de propriedade (TCO) da AWS)
- Categorias de armazenamento do Amazon S3
- Produtos da Nuvem AWS
- Modelo de responsabilidade compartilhada da AWS

Vídeos relacionados:

- Why move to a managed database? (Por que migrar para um banco de dados gerenciado?)
- What is Amazon EMR and how can I use it for processing data? (O que é o Amazon EMR e como usá-lo para processar dados?)

Exemplos relacionados:

- Why to move to a managed database (Por que migrar para um banco de dados gerenciado?)
- Consolidate data from identical SQL Server databases into a single Amazon RDS for SQL Server database using AWS DMS (Consolidar dados de bancos de dados idênticos do Consolidate SQL Server em um banco de dados único do Amazon RDS for SQL Server usando o AWS DMS)
- Deliver data at scale to Amazon Managed Streaming for Apache Kafka (Amazon MSK) (Entregar dados em grande escala para o Amazon Managed Streaming for Apache Kafka (Amazon MSK))
- Migrate an ASP.NET web application to AWS Elastic Beanstalk (Migrar uma aplicação Web ASP.NET para o AWS Elastic Beanstalk)

COST05-BP04 Selecionar software com licenciamento econômico

Os softwares de código aberto eliminam 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 exposição a riscos se esta prática recomendada não for estabelecida: baixo

Orientações para a implementação

O código aberto originou-se no contexto do desenvolvimento de software para indicar que o software está em conformidade com determinados critérios de distribuição gratuita. O software de código aberto é composto de código-fonte que pode ser inspecionado, modificado e aprimorado por qualquer pessoa. Com base nos requisitos de negócios, nas habilidades dos engenheiros, no uso previsto ou em outras dependências tecnológicas, as organizações podem considerar o uso de software de código aberto na AWS para minimizar os custos de licença. Ou seja, o custo das licenças de software pode ser reduzido 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.

Avalie os benefícios do software licenciado em relação ao custo total para otimizar a workload. 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.

Por exemplo, operar uma instância do Amazon EC2 na us-east-1 com um sistema operacional Linux permite reduzir os custos em aproximadamente 45%, em comparação com a execução de outra instância do Amazon EC2 no Windows.

O AWS Pricing Calculator oferece uma maneira abrangente de comparar os custos de vários recursos com diferentes opções de licença, como as instâncias do Amazon RDS e diferentes mecanismos de banco de dados. Além disso, o AWS Cost Explorer fornece uma perspectiva inestimável dos custos das workloads existentes, especialmente daquelas que vêm com licenças diferentes. Para gerenciamento de licenças, o AWS License Manager oferece um método simplificado para supervisionamento e gerenciamento de licenças de software. Os clientes podem implantar e operacionalizar o software de código aberto preferido na Nuvem AWS.

Etapas da implementação

- Analisar as 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 fornecedor do software: analise todo o histórico de alterações de preços ou de 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, verifique como eles executam auditorias e as penalidades que poderiam ser impostas.

Recursos

Documentos relacionados:

- Open Source at AWS
- Calculadora de custo total de propriedade (TCO) da AWS
- Categorias de armazenamento do Amazon S3
- Produtos da nuvem

Exemplos relacionados:

- Blogs de código aberto
- Blogs de código aberto da AWS
- Otimização e Avaliação de Licenciamento

COST05-BP05 Selecionar os componentes desta workload para otimizar o custo alinhado com as prioridades da organização

Considere o custo ao selecionar todos os componentes para sua workload. Isso inclui o uso de serviços gerenciados e em nível de aplicação ou arquitetura sem servidor, contêineres ou orientada a eventos a fim de reduzir o custo geral. Minimize os custos de licença usando um software de código aberto ou que não tenha taxas de licença ou alternativas para reduzir os gastos.

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

Orientação para implementação

Considere o custo de serviços e opções 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, o Amazon Simple Notification Service (Amazon SNS) e o 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 e o Amazon Simple Storage Service (Amazon S3) para sites estáticos. Se possível, coloque sua aplicação em contêineres e use serviços de contêiner gerenciados pela AWS, como o Amazon Elastic Container Service (Amazon ECS) ou Amazon Elastic Kubernetes Service (Amazon EKS).

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).

você pode usar serviços sem servidor ou em nível de aplicativo, como o Lambda, o Amazon Simple Queue Service (Amazon SQS), Amazon SNSe o 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.

O uso de arquitetura orientada a eventos também é possível com serviços sem servidor. Arquiteturas orientadas a eventos são baseadas em push, então, tudo acontece sob demanda à medida que o evento se apresenta no roteador. Dessa forma, você não paga pela sondagem contínua para conferir um evento. Isso significa um consumo menor de largura de banda de rede, menor utilização de CPU, menor capacidade de frota ociosa e menos handshakes SSL/TLS.

Para obter mais informações sobre tecnologia sem servidor, consulte 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. Em vez de aumentar a capacidade para atender à demanda, considere outras opções que podem oferecer melhor performance por um custo menor. Por exemplo, se você precisar analisar o tráfego esperado para seus bancos de dados na AWS e pensar em aumentar o tamanho da instância ou usar serviços do Amazon ElastiCache (Redis ou Memcached) a fim de fornecer mecanismos em cache para seus bancos de dados.
- Avaliar a arquitetura orientada a eventos: O uso de uma arquitetura sem servidor também permite criar uma arquitetura orientada a eventos para aplicações distribuídas e baseadas em microsserviço, o que ajuda a criar soluções escaláveis, resilientes, ágeis e econômicas.

Recursos

Documentos relacionados:

- Calculadora de preços da AWS
- tecnologia sem servidor da AWS
- O que é arquitetura orientada a eventos
- Categorias de armazenamento do Amazon S3
- Produtos da nuvem
- Amazon ElastiCache for Redis

Exemplos relacionados:

- Comece a usar a arquitetura orientada a eventos
- Arquitetura orientada a eventos
- How Statsig runs 100x more cost-effectively using Amazon ElastiCache for Redis
- Best practices for working with AWS Lambda functions

COST05-BP06 Realizar análises de custos para diferentes usos ao longo do tempo

As workloads 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 exposição a riscos quando esta prática recomendada não é estabelecida: Médio

Orientações para a 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 workload depende dos requisitos da sua organização. Se for uma workload 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.

Se precisar mover dados para a Nuvem AWS, você poderá selecionar qualquer série de serviços que a AWS oferece e ferramentas de parceiros para ajudar a migrar seus conjuntos de dados, sejam

eles arquivos, bancos de dados, imagens de máquina, volumes de bloco ou até backups de fita. Por exemplo, para mover um grande volume de dados para a AWS e dela ou processar dados na borda, você pode usar um dos dispositivos com propósito específico da AWS para mover petabytes de dados offline de forma econômica. Outro exemplo é relativo a taxas de transferência de dados mais altas, um serviço de conexão direta pode ser mais barato do que uma VPN, que fornece a conectividade consistente necessária para sua empresa.

Com base na análise de custos para uso diferente no decorrer do tempo, analise sua atividade de escalabilidade. Analise o resultado para ver se a política de escalabilidade pode ser ajustada para adicionar instâncias de vários tipos e opções de compra. Analise suas configurações para verificar se é possível reduzir o mínimo para atender às solicitações do usuário, mas com um tamanho de frota menor e adicionar mais recursos para atender à alta demanda esperada.

Realize uma análise de custo para uso diferente no decorrer do tempo conversando com os stakeholders em sua organização e use o recurso de previsão do AWS Cost Explorer para prever o possível impacto das alterações de serviço. Monitore os gatilhos de nível de uso utilizando o AWS Budgets, alarmes de faturamento do CloudWatch e o AWS Cost Anomaly Detection para identificar e implementar os serviços mais econômicos com maior rapidez.

Etapas da implementação

- Definição de 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. Converse com os stakeholders da empresa sobre aumentos de uso e custos históricos e previstos e garanta que os aumentos se alinhem com os requisitos da empresa. Identifique os dias, as semanas ou os meses em que você espera que mais usuários utilizem seus recursos da AWS, o que indica que você deve aumentar a capacidade dos recursos existentes ou adotar serviços adicionais a fim de reduzir o custo e aumentar a performance.
- Realize a análise de custos e uso previsto: usando os padrões de uso definidos, realize 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. Em outras palavras, quando o custo aumenta, o uso também deve aumentar para a empresa.

Recursos

Documentos relacionados:

- AWS Total Cost of Ownership (TCO) Calculator (Calculadora de custo total de propriedade (TCO) da AWS)
- Categorias de armazenamento do Amazon S3
- Produtos da nuvem
- Amazon EC2 Auto Scaling
- Migração de dados para a nuvem
- AWS Snow Family

Vídeos relacionados:

- AWS OpsHub for Snow Family

CUSTOS 6. Como atingir 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 Realizar modelagem de custos
- COST06-BP02 Selecionar o tipo, o tamanho e o número do recurso com base nos dados
- COST06-BP03 Selecionar o tipo, tamanho e número do recurso automaticamente com base nas métricas

COST06-BP01 Realizar modelagem de custos

Identifique os requisitos da organização (como as necessidades dos negócios e os compromissos existentes) e realize a modelagem dos custos (custos gerais) da workload e de cada um de seus componentes. Realize atividades de referência para a workload 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 é estabelecida: alto

Orientações para a implementação

Execute a modelagem de custos para sua workload e cada um de seus componentes para entender o equilíbrio entre recursos e encontrar o tamanho correto para cada recurso na workload, considerando um nível específico de performance. O entendimento das considerações de custo pode embasar seu processo de tomada de decisão e caso de negócios organizacional ao avaliar os resultados da realização de valor para a implantação planejada da workload.

Realize atividades de referência para a workload 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 saber as práticas recomendadas, consulte a seção Review do Performance Efficiency Pillar of the AWS Well-Architected Framework.

Por exemplo, para criar a modelagem de custos para uma workload que consista em recursos de computação, o AWS Compute Optimizer pode auxiliar com a modelagem de custos para executar workloads. Ele fornece recomendações de dimensionamento correto para recursos de computação com base no uso histórico. Implante os CloudWatch Agents nas instâncias do Amazon EC2 para coletar métricas de memória que ajudam você com recomendações mais precisas no AWS Compute Optimizer. 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.

Há vários serviços que você pode usar com logs personalizados como fontes de dados para dimensionar adequadamente as operações para outros serviços e componentes da workload, como o AWS Trusted Advisor, o Amazon CloudWatch e o Amazon CloudWatch Logs. O AWS Trusted Advisor confere os recursos e sinaliza os que apresentam baixa utilização, o que pode ajudar você a dimensioná-los corretamente e criar uma modelagem de custo.

Veja a seguir as recomendações para dados e métricas de modelagem de custo:

- O monitoramento deve refletir com precisão a experiência do usuário. 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 workload. 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.
- Escolha os serviços da AWS certos para sua workload planejada considerando seus compromissos existentes, modelos de preço selecionados para outras workloads e a capacidade de inovar com maior rapidez e concentrar-se em seu valor comercial principal.

Etapas da implementação

- Realização de modelagem de custo para os recursos: 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, junto com os dados de custo da hora em que o teste foi executado. Depois, reimplante a workload ou altere os tipos e tamanhos de recursos e execute novamente o teste. Inclua taxas de licença para todos os produtos que você pode usar com esses recursos e custos de operações estimados (mão de obra ou engenharia) para implantar e gerenciar esses recursos ao criar a modelagem de custo. Considere a modelagem de custo para um período (por hora, diária, anual ou três anos).

Recursos

Documentos relacionados:

- AWS Auto Scaling
- Identifying Opportunities to Right Size (Identificar oportunidades para dimensionar o tamanho corretamente)
- Amazon CloudWatch features (Recursos do Amazon CloudWatch)
- Cost Optimization: Amazon EC2 Right Sizing (Otimização de custos: dimensionamento correto do Amazon EC2)
- AWS Compute Optimizer
- AWS Pricing Calculator (Calculadora de preços da AWS)

Exemplos relacionados:

- Perform a Data-Driven Cost Modelling (Executar uma modelagem de custo orientada a dados)
- Estimate the cost of planned AWS resource configurations (Estimar o custo das configurações de recursos planejados da AWS)
- Choose the right AWS tools (Selecionar as ferramentas certas da AWS)

COST06-BP02 Selecionar o tipo, o tamanho e o número do recurso com base nos dados

Selecione o tamanho ou tipo do 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, a documentação ou outras fontes de informações sobre a workload.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

O Amazon EC2 fornece uma ampla seleção de tipos de instância com diferentes níveis de capacidade de CPU, memória, armazenamento e rede para atender a diferentes casos de uso. Esses tipos de instância dispõem de diferentes combinações de capacidade de CPU, memória, armazenamento e rede, oferecendo versatilidade ao selecionar a combinação certa de recursos para os projetos. Eles são disponibilizados em vários tamanhos para que seja possível ajustar os recursos com base nas demandas da workload. Para determinar o tipo de instância necessário, reúna os detalhes dos requisitos do sistema da aplicação ou do software a ser executado na instância. Esses detalhes devem incluir:

- Sistema operacional
- Número de núcleos de CPU
- Núcleos de GPU
- Quantidade de memória do sistema (RAM)
- Tipo e espaço de armazenamento
- Requisito de largura de banda da rede

Identifique a finalidade dos requisitos de computação e a instância necessária e conheça as várias famílias de instâncias do Amazon EC2. A Amazon oferece as seguintes famílias de tipos de instância:

- De uso geral
- Otimizadas para computação
- Otimizadas para memória
- Otimizadas para armazenamento
- Computação acelerada
- Otimizadas para HPC

Para compreender melhor os propósitos e casos de uso específicos aos quais determinada família de instâncias do Amazon EC2 pode atender, consulte Tipos de instância da AWS.

A coleta dos requisitos do sistema é essencial para selecionar a família e o tipo de instância específicos que melhor atendem às suas necessidades. Os nomes dos tipos de instância são compostos do nome da família e do tamanho da instância. Por exemplo, a instância t2.micro é da família T2 e é de tamanho micro.

Selecione o tamanho ou o 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 custos, uma versão anterior da workload (como uma versão on-premises), a documentação ou outras fontes de informações sobre a workload (whitepapers ou soluções publicadas). O uso de calculadoras de preços ou de ferramentas de gerenciamento de custos da AWS pode ajudar a tomar decisões fundamentadas sobre tipos, tamanhos e configurações de instância.

Etapas da implementação

- Selecionar recursos com base nos dados: use os dados da modelagem de custos para selecionar o nível de uso previsto da workload e escolher o tipo e o tamanho de recurso especificados. Com base nos dados da modelagem de custos, determine o número de CPUs virtuais, a memória total (GiB), o volume de armazenamento de instâncias local (GB), os volumes do Amazon EBS e o nível de desempenho da rede, levando em consideração a taxa de transferência de dados necessária para a instância. Sempre faça seleções com base em análise detalhada e em dados precisos para otimizar o desempenho e, ao mesmo tempo, gerenciar os custos de forma eficiente.

Recursos

Documentos relacionados:

- Tipos de instância da AWS
- AWS Auto Scaling
- Amazon CloudWatch features (Recursos do Amazon CloudWatch)
- Otimização de custos: dimensionamento correto do EC2

Vídeos relacionados:

- Selecting the right Amazon EC2 instance for your workloads
- Right size your service

Exemplos relacionados:

- It just got easier to discover and compare Amazon EC2 instance types

COST06-BP03 Selecionar o tipo, tamanho e número do recurso automaticamente com base nas métricas

Use métricas da workload 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 de computação, armazenamento, dados e rede. Isso pode ser feito com um ciclo de comentários, como escalabilidade automática ou por código personalizado na workload.

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

Orientações para a implementação

Crie um loop de comentários dentro da workload que usa métricas ativas da workload em execução para fazer alterações nessa workload. Você pode usar um serviço gerenciado, como AWS Auto Scaling, que você configura para executar as operações de dimensionamento corretas para você. O AWS também fornece APIs, SDKs e 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 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 do AWS possuem seleção automática de tipo ou tamanho, como Amazon Simple Storage Service 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

- Aumentar sua observabilidade configurando métricas de workload: Capture as principais métricas para a workload. Essas métricas fornecem uma indicação da experiência do cliente, como a saída da workload, e se alinham às diferenças entre tipos e tamanhos de recursos, como uso de CPU e memória. Para recursos de computação, analise os dados de desempenho para dimensionar corretamente suas instâncias do Amazon EC2. Identifique instâncias ociosas e subutilizadas. As principais métricas a serem procuradas são o uso da CPU e a utilização da memória (por exemplo, 40% de utilização da CPU em 90% do tempo, conforme explicado em Dimensionamento correto

com o AWS Compute Optimizer e utilização da memória ativada). Identifique instâncias com uso máximo de CPU e utilização de memória inferior a 40% em um período de quatro semanas. São as instâncias no tamanho certo para reduzir custos. Para recursos de armazenamento, como Amazon S3, você pode usar a Lente de Armazenamento do Amazon S3, que permite ver 28 métricas em várias categorias no nível do bucket e 14 dias de dados históricos no painel por padrão. Você pode filtrar seu painel da Lente de Armazenamento do Amazon S3 por resumo e otimização de custos ou eventos para analisar métricas específicas.

- Ver recomendações de redimensionamento: Use as recomendações de dimensionamento correto no AWS Compute Optimizer e a ferramenta de dimensionamento correto do Amazon EC2 no console de gerenciamento de custos ou revise o dimensionamento correto do AWS Trusted Advisor de seus recursos para fazer ajustes em sua workload. É importante usar as ferramentas certas ao dimensionar diferentes recursos e seguir as diretrizes de dimensionamento correto, sejam instâncias do Amazon EC2, classes de armazenamento do AWS ou tipos de instância do Amazon RDS. Para recursos de armazenamento, é possível usar a Lente de Armazenamento do Amazon S3, que oferece visibilidade do uso de armazenamento de objetos e tendências de atividade, bem como faz recomendações acionáveis para otimizar custos e aplicar as práticas recomendadas de proteção de dados. Usando as recomendações contextuais que a Lente de Armazenamento do Amazon S3 obtém da análise de métricas em sua organização, você pode tomar medidas imediatas para otimizar seu armazenamento.
- Selecionar o tipo e o tamanho do recurso automaticamente com base nas métricas: Usando as métricas de workload, selecione manual ou automaticamente seus recursos de workload. Para recursos de computação, 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. Você pode iniciar e dimensionar automaticamente uma frota de instâncias sob demanda e instâncias spot em um único grupo do Auto Scaling. Além de receber descontos pelo uso de instâncias spot, você pode usar instâncias reservadas ou um Savings Plan para receber taxas com desconto do preço regular da instância sob demanda. Todos esses fatores combinados ajudam você a otimizar sua economia de custos para instâncias do Amazon EC2 e determinar a escala e o desempenho desejados para seu aplicativo. Você também pode usar uma estratégia de seleção de tipo de instância baseada em atributo (ABS) em Grupos do Auto Scaling (ASG), que permite expressar seus requisitos de instância como um conjunto de atributos, como vCPU, memória e armazenamento. Você pode usar automaticamente os tipos de instância de geração mais recente quando eles são lançados e acessar uma variedade mais ampla de capacidade com instâncias spot do Amazon EC2. A frota do Amazon EC2 e o Amazon EC2 Auto Scaling selecionam e executam instâncias que se ajustam aos atributos especificados, eliminando a necessidade de escolher manualmente os tipos

de instância. Para recursos de armazenamento, você pode usar os recursos Amazon S3 Intelligent Tiering e Amazon EFS Infrequent Access, que permitem selecionar classes de armazenamento automaticamente que oferecem economia automática de custos de armazenamento quando os padrões de acesso aos dados mudam, sem impacto no desempenho ou sobrecarga operacional.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- Dimensionamento correto do AWS
- AWS Compute Optimizer
- Recursos do Amazon CloudWatch
- Configuração do CloudWatch
- Publicar métricas personalizadas do CloudWatch
- Conceitos básicos do Amazon EC2 Auto Scaling
- Lente de Armazenamento do Amazon S3
- Amazon S3 Intelligent-Tiering
- Amazon EFS Infrequent Access
- Iniciar uma instância do Amazon EC2 usando o SDK

Vídeos relacionados:

- Dimensionar corretamente seus serviços

Exemplos relacionados:

- Seleção de tipo de instância baseada em atributo do Auto Scaling para a frota do Amazon EC2
- Otimização do Amazon Elastic Container Service para o custo usando escalabilidade programada
- Escalabilidade preditiva com o Amazon EC2 Auto Scaling
- Otimizar os custos e ganhar visibilidade no uso com a Lente de Armazenamento do Amazon S3
- Laboratórios do Well-Architected: Recomendações de dimensionamento correto (Nível 100)
- Laboratórios do Well-Architected: Dimensionamento correto com o AWS Compute Optimizer e utilização de memória habilitada (Nível 200)

CUSTOS 7. Como usar os 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 preço
- COST07-BP02 Escolher regiões com base no custo
- COST07-BP03 Selecionar contratos de terceiros com termos econômicos
- COST07-BP04 Implementar modelos de preços para todos os componentes dessa workload
- COST07-BP05 Realizar análise de modelo de preço em nível da conta de gerenciamento

COST07-BP01 Executar análise de modelo de preço

Analise cada componente da workload. Determine se o componente e os recursos serão executados por períodos estendidos (para descontos de compromisso) ou dinâmicos e curtos (para spot ou sob demanda). Execute uma análise da workload usando as recomendações nas ferramentas de gerenciamento de custos e aplique regras de negócios a essas recomendações para alcançar altos retornos.

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

Orientações para a implementação

A AWS tem vários modelos de preço que permitem que você pague pelos seus recursos da maneira mais econômica que atenda às necessidades de sua organização e de acordo com o produto. Trabalhe com suas equipes para determinar o modelo de preço mais apropriado. Com frequência, o modelo de preço consiste em uma combinação de várias opções, tal como determinado por seus requisitos de disponibilidade.

As instâncias sob demanda permitem que você pague por capacidade computacional ou de banco de dados por hora ou por segundo (60 segundos, no mínimo), dependendo de quais instâncias são executadas, sem compromissos de longo prazo nem pagamentos adiantados.

Os Savings Plans são um modelo de preço flexível que oferece preços baixos para uso do Amazon EC2, do Lambda e do AWS Fargate (Fargate), em troca do compromisso com uma quantidade de uso consistente (medida em dólares por hora) por um período de um ou três anos.

As instâncias spot são um mecanismo de preço do Amazon EC2 que permite que você solicite capacidade computacional extra por uma taxa por hora com desconto (até 90% de desconto no preço sob demanda) sem compromisso inicial.

As instância reservadas permitem que você obtenha um desconto de até 75% com o pagamento antecipado de capacidade. Para obter mais detalhes, consulte Otimização de custos por meio de reservas.

Você pode optar por incluir um Savings Plan para os recursos associados aos ambientes de produção, qualidade e desenvolvimento. Visto que os recursos da área restrita para testes são fornecidos somente quando necessários, você também pode optar por um modelo sob demanda para os recursos nesse ambiente. Use instâncias spot da Amazon para reduzir os Amazon EC2 custos ou use Savings Plans para computação para reduzir o custo do Amazon EC2, do Fargate e do Lambda. A ferramenta de recomendações AWS Cost Explorer oferece oportunidades de descontos de compromisso com o Saving Plans.

Se alguma vez você já comprou instâncias reservadas para o Amazon EC2 ou estabeleceu práticas de alocação de custos em sua organização, poderá continuar usando as instâncias reservadas do Amazon EC2 por enquanto. Entretanto, recomendamos elaborar uma estratégia para usar Savings Plans no futuro como um mecanismo de redução de custos mais flexível. Você pode atualizar as recomendações de Savings Plans (SP) no AWS Cost Management para gerar novas recomendações de Savings Plans sempre que quiser. Use instâncias reservadas (IR) para reduzir os custos do Amazon RDS, do Amazon Redshift, do Amazon ElastiCache e do Amazon OpenSearch Service. Os Saving Plans e as instâncias reservadas estão disponíveis em três opções: pagamento adiantado, pagamento adiantado parcial e sem pagamento adiantado. Use as recomendações de compra de IR e SP fornecidas no AWS Cost Explorer.

Para encontrar oportunidades para workloads spot, use uma visualização por hora do uso geral e procure períodos regulares de uso ou elasticidade variáveis. Você pode usar instâncias spot para diversas aplicações tolerantes a falhas e flexíveis. Exemplos incluem servidores Web sem estado, endpoints de API, aplicações de big data e análise, workloads conteinerizadas, CI/CD e outras workloads flexíveis.

Analise suas instâncias do Amazon EC2 e do Amazon RDS para ver se elas podem ser desativadas quando não estiverem em uso (após o expediente e nos fins de semana). Essa abordagem permitirá que você reduza os custos em 70% ou mais em comparação a usá-las ininterruptamente. Se você tiver clusters do Amazon Redshift necessários apenas em momentos específicos, poderá pausar o cluster e, posteriormente, retomá-lo. Quando se interrompe o cluster do Amazon Redshift ou

a instância do Amazon EC2 e do Amazon RDS, o faturamento de computação é interrompido e somente se aplica a cobrança de armazenamento.

Observe que as reservas de capacidade sob demanda (ODCR) não são um desconto de preço. As reservas de capacidade são cobradas pela taxa sob demanda equivalente, quer você execute ou não as instâncias na capacidade reservada. Elas devem ser consideradas quando você precisa fornecer capacidade suficiente para os recursos que pretende executar. As ODCRs não precisam estar atreladas a compromissos de longo prazo, visto que elas podem ser canceladas quando não mais necessárias, mas elas também podem se beneficiar dos descontos que os Savings Plans ou as instâncias reservadas oferecem.

Etapas da implementação

- Análise da 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
- Análise dos contratos de preço existentes: examine contratos ou compromissos atuais para necessidades de longo prazo. Analise o que você tem no momento e quanto esses compromissos estão em uso. Utilize os descontos contratuais ou contratos empresariais preexistentes. Os contratos empresariais oferecem aos clientes a opção de personalizar acordos que melhor atendam às necessidades deles. Com relação a compromissos de longo prazo, considere descontos de preço reservados, instâncias reservadas ou Savings Plans para o tipo específico de instância, a família de instâncias, a Região da AWS e as zonas de disponibilidade.
- Realização de uma análise de descontos de compromisso: usando o Cost Explorer em sua conta, examine as recomendações de Savings Plans e instâncias reservadas. Para garantir que você implemente as recomendações corretas com os descontos e riscos necessários, siga os laboratórios do Well-Architected.

Recursos

Documentos relacionados:

- Accessing Reserved Instance recommendations (Como acessar as recomendações de instâncias reservadas)

- Opções de compra de instância
- AWS Enterprise

Vídeos relacionados:

- Economize até 90% e execute workloads 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 preço

COST07-BP02 Escolher regiões com base no custo

A definição de preço dos recursos pode ser diferente em cada região. Identifique as diferenças de custo regionais e implante apenas nas regiões com custos mais altos para atender aos requisitos de latência, residência e soberania de dados. A consideração do custo da região ajuda você a pagar o menor preço geral por essa workload.

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

Orientação para implementação

A infraestrutura da Nuvem AWS é global, hospedada em vários locais em todo o mundoe criada em torno de Regiões da AWS, zonas de disponibilidade, zonas locais, AWS Outposts e zonas do Wavelength. Uma região é um local físico no mundo, e cada região é uma área geográfica separada onde a AWS tem várias zonas de disponibilidade. As zonas de disponibilidade, que são locais isolados em cada região, consistem em um ou mais datacenters discretos, cada um com energia, rede e conectividade redundantes.

Cada Região da AWS opera de acordo com as condições do mercado local, e a definição de preço dos recursos é diferente em cada região devido às diferenças de custo de terra, fibra, eletricidade e impostos, por exemplo. Escolha uma região específica para operar um componente de sua solução completa para que você possa operar ao menor preço possível globalmente. Use a Calculadora da AWS para calcular os custos de sua workload em várias regiões procurando serviços por tipo de local (região, zona do Wavelength e zona local) e região.

Ao projetar suas soluções, uma prática recomendada é buscar colocar os recursos de computação mais perto dos usuários para proporcionar menor latência e forte soberania de dados. Selecione a localização geográfica com base nos requisitos de segurança, performance, privacidade de dados e empresariais. Para aplicações com usuários finais globais, use várias localidades.

Use regiões que ofereçam preços mais baixos por serviços da AWS para implantar suas workloads se você não tiver obrigações em requisitos de privacidade de dados, segurança e empresariais. Por exemplo, se sua região padrão for ap-southeasth-2 (Sydney) e não houver restrições (privacidade de dados, segurança, por exemplo) quanto ao uso de outras regiões, a implantação de instâncias não essenciais (desenvolvimento e teste) Amazon EC2 na região north-east-1 (N. da Virgínia) custará menos.

	Conformidade	*Latência*	*Custos*	*Serviços/recursos*
Região 1	✓	15 ms	$$	✓
Região 2	✓	20 ms	$$$	X
Região 3	✓	80 ms	$	✓
Região 4	✓	15 ms	$$	✓
Região 5	✓	20 ms	$$$	X
Região 6	✓	15 ms	$	✓
Região 7	✓	80 ms	$	✓
Região 8	✓	15 ms	$	X

Tabela de matriz de recursos da região

A tabela de matriz anterior mostra que a região 4 é a melhor opção para esse determinado cenário porque a latência é baixa em comparação a outras regiões, o serviço está disponível e é a região mais barata.

Etapas da implementação

- Revise a definição de preço da Região da AWS: 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.

- Analise os requisitos para implantações em várias regiões: analise seus requisitos e obrigações empresariais (privacidade de dados, segurança ou performance) para descobrir se há restrições quanto ao uso de vários regiões. Se não houver obrigações que restrinjam você ao uso de uma região, use várias regiões.
- Analise a transferência de dados necessária: Leve em conta os custos de transferência de dados ao selecionar regiões. Mantenha seus dados perto de seu cliente e dos recursos. Selecione Regiões da AWS mais baratas onde os dados fluam e haja transferência de dados mínima. Dependendo dos requisitos empresariais para transferência de dados, você pode usar Amazon CloudFront, o AWS PrivateLink, o AWS Direct Connecte o AWS Virtual Private Network para reduzir seus custos de rede, melhorar a performance e aprimorar a segurança.

Recursos

Documentos relacionados:

- Acesso a recomendações de instância reservada
- Definição de preço do Amazon EC2
- Opções de compra de instância
- Region Table (Tabela de regiões)

Vídeos relacionados:

- Economize até 90% e execute cargas de trabalho de produção no local

Exemplos relacionados:

- Overview of Data Transfer Costs for Common Architectures (Visão geral dos custos de transferência de dados para arquiteturas comuns)
- Cost Considerations for Global Deployments (Considerações de custo para implantações globais)
- O que considerar ao selecionar uma região para suas workloads
- Well-Architected Labs: Restrict service usage by Region (Level 200) (Well-Architected Labs: uso restrito do serviço por região (nível 200))

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 se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

Existem vários produtos no mercado que podem ajudar você a gerenciar os custos em ambientes de nuvem. Eles podem ter algumas diferenças em termos de recursos que dependem dos requisitos do cliente, como alguns que enfatizam a governança ou a visibilidade dos custos e outros a otimização de custos. Um fator-chave para a eficácia da otimização e da governança de custos é usar a ferramenta certa com os recursos necessários e o modelo de preços correto. Esses produtos têm modelos de preços diferentes. Alguns aplicam determinada porcentagem de cobrança sobre sua fatura mensal, enquanto outros aplicam uma porcentagem sobre as economias obtidas. O ideal é pagar apenas pelo que você precisa.

Ao usar soluções ou serviços de terceiros na nuvem, é importante que as estruturas de preços estejam alinhadas aos resultados desejados. A definição de preço deve ser dimensionada de acordo com os resultados e o valor que fornece. Por exemplo, em software que leva uma porcentagem das economias que ele fornece, quanto mais você economiza (resultado), mais ele cobra. Os contratos de licença em que você paga mais conforme suas despesas aumentam nem sempre podem ser vantajosos em termos de otimização de custos. No entanto, se o fornecedor oferecer benefícios claros para todos os componentes da sua fatura, talvez essa taxa de ajuste de escala seja aceitável.

Por exemplo, uma solução que fornece recomendações para o Amazon EC2 e que aplica uma porcentagem de cobrança sobre toda a fatura poderá se tornar mais cara se você usar outros serviços que não oferecem nenhum benefício. Outro exemplo é um serviço gerenciado que é cobrado segundo uma porcentagem do custo dos recursos gerenciados. O tamanho maior de uma instância pode não exigir necessariamente maior esforço de gerenciamento, mas pode ter uma cobrança maior. Verifique se essas disposições de definição de preços de serviços incluem um programa ou recursos de otimização de custos no respectivo serviço para promover a eficiência.

Os clientes podem encontrar produtos mais avançados ou mais fáceis de usar no mercado. Você precisa considerar o custo desses produtos e avaliar possíveis resultados da otimização de custos a longo prazo.

Etapas da implementação

- Analisar contratos e termos de terceiros: analise os preços em contratos com terceiros. Execute modelagem para diferentes níveis de uso e leve em consideração novos custos, como o 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:

- Accessing Reserved Instance recommendations (Como acessar as recomendações de instâncias reservadas)
- Opções de compra de instância

Vídeos relacionados:

- Economize até 90% e execute workloads de produção no local

COST07-BP04 Implementar modelos de preços 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 são executadas por tempo suficiente para a capacidade reservada, entre 25% e 75% do período, dependendo do tipo do recurso.

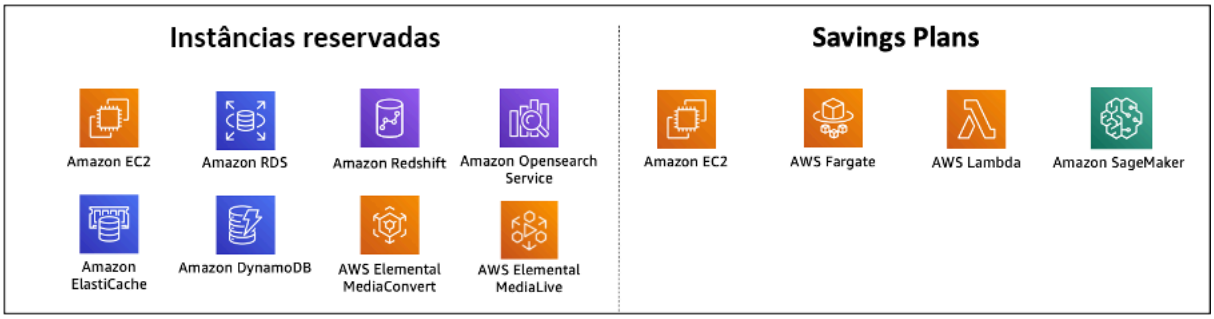
Nível de exposição a riscos se esta prática recomendada não for estabelecida: baixo

Orientações para a implementação

Para melhorar o custo-benefício, a AWS fornece várias recomendações de compromisso com base no uso anterior. Essas recomendações podem ser usadas para compreender o que você pode economizar e como o compromisso será usado. É possível usar esses serviços como instâncias sob demanda ou spot ou assumir um compromisso por determinado período e reduzir os custos sob demanda com instâncias reservadas (RIs) e Savings Plans (SPs). É necessário compreender, além de cada componente da workload e dos vários serviços da AWS, os descontos de compromisso, as opções de compra e as instâncias spot desses serviços para otimizar a workload.

Considere os requisitos dos componentes da workload e informe-se sobre os diferentes modelos de preços desses serviços. Defina o requisito de disponibilidade desses componentes. 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.

Por exemplo, vamos analisar essa arquitetura de aplicações web na AWS. Esse exemplo de workload consiste em vários serviços da AWS, como o Amazon Route 53, o AWS WAF, o Amazon CloudFront, as instâncias do Amazon EC2, as instâncias do Amazon RDS, os balanceadores de carga, o armazenamento do Amazon S3 e o Amazon Elastic File System (Amazon EFS). Você precisa analisar cada um desses serviços e identificar as possíveis oportunidades de redução de custos com diferentes modelos de preços. Alguns deles podem ser elegíveis para IRs ou SPs, e outros podem estar disponíveis apenas sob demanda. Como mostrado na imagem a seguir, alguns dos serviços da AWS podem ser compromissados usando IRs ou SPs.



Serviços da AWS compromissados que usam instâncias reservadas e Savings Plans

Etapas da implementação

- Implementar modelos de preços: usando os resultados da análise, compre Savings Plans, instâncias reservadas ou implemente instâncias spot. Se esta for a sua primeira compra de compromisso, escolha as cinco ou dez principais recomendações da lista, monitore e analise os resultados de um ou dos dois próximos meses. O AWS Cost Management Console fornece orientações durante o processo. Analise as recomendações de IR ou de SP no console, personalize as recomendações (tipo, pagamento e prazo) e analise o compromisso por hora (por exemplo, USD 20 por hora) e adicione ao carrinho. Os descontos se aplicam automaticamente ao uso qualificado. Compre uma pequena quantidade 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 que são sem estado. Por fim, selecione instâncias sob demanda do Amazon EC2 e aloque recursos para os demais requisitos.

- Ciclo de análise da workload: implemente um ciclo de análise da workload que examine especificamente a cobertura do modelo de preços. Assim que a workload tiver a cobertura necessária, compre descontos de compromisso adicionais parcialmente (a cada dois meses) ou conforme o uso da sua organização mudar.

Recursos

Documentos relacionados:

- Understanding your Savings Plans recommendations
- Acesso a recomendações de instância reservada
- Como comprar instâncias reservadas
- Opções de compra de instância
- Instâncias spot
- Modelos de reserva para outros serviços da AWS
- Savings Plans Supported Services

Vídeos relacionados:

- Economize até 90% e execute workloads de produção no local

Exemplos relacionados:

- What should you consider before purchasing Savings Plans?
- How can I use Cost Explorer to analyze my spending and usage?

COST07-BP05 Realizar análise de modelo de preço em nível da conta de gerenciamento

Confira as ferramentas de gerenciamento de faturamento e de custos e veja os descontos recomendados com compromissos e reservas para realizar uma análise regular no nível da conta de gerenciamento.

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

Orientação para implementação

A execução de uma modelagem de custo regular ajuda você a implementar oportunidades de otimização em várias workloads. 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 poderá atingir 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 preço continue a evoluir com suas workloads dinâmicas e os respectivos componentes.

Use a ferramenta de recomendações do AWS Cost Explorer para encontrar oportunidades de descontos de compromisso em sua conta de gerenciamento. As recomendações em nível de conta de gerenciamento são calculadas considerando-se o uso em todas as contas da organização da AWS que têm instâncias reservadas (RI) ou Savings Plans (SP). Elas também são calculadas quando o compartilhamento de descontos é ativado para recomendar um compromisso que maximize a economia em todas as contas.

Embora a compra em nível da conta de gerenciamento seja otimizada para obter o máximo de economia em muitos casos, pode haver situações em que você considere comprar SPs em nível da conta vinculada, como quando você deseja que os descontos se apliquem primeiro ao uso nessa conta vinculada específica. As recomendações da conta principal são calculadas em nível de conta individual para maximizar as economias em cada conta isolada. Se sua conta tiver compromissos de RI e SP, eles serão aplicados na seguinte ordem:

1. RI de zona
2. RI padrão
3. RI conversível
4. Plano de economia de instâncias
5. Plano de economia de computação

Se você comprar um SP em nível da conta de gerenciamento, a economia será aplicada com base na porcentagem de desconto mais alta para a mais baixa. Os SPs em nível da conta de gerenciamento examinam todas as contas vinculadas e aplicarão as economias sempre que o desconto for maior. Se desejar restringir onde as economias são aplicadas, você pode comprar um Savings Plan em nível da conta vinculada e, sempre que a conta estiver executando serviços computacionais qualificados, o desconto será aplicado primeiro. Quando a conta não estiver executando serviços computacionais qualificados, o desconto será compartilhado entre as outras

contas vinculadas na mesma conta de gerenciamento. O compartilhamento de descontos está ativado por padrão, mas pode ser desativado se necessário.

Em uma família de faturamento consolidado, os Savings Plans são aplicados primeiro ao uso da conta do proprietário e depois ao uso de outras contas. Isso ocorrerá somente se você tiver o compartilhamento habilitado. Seus Savings Plans são aplicados primeiro à sua maior porcentagem de economia. Se houver vários usos com porcentagens de economia iguais, Savings Plans serão aplicados ao primeiro uso com a taxa mais baixa de Savings Plans. Os Savings Plans continuam a ser aplicados até que não haja mais usos restantes ou que seu compromisso seja esgotado. Qualquer uso restante é cobrado de acordo com as tarifas sob demanda. É possível atualizar as recomendações de Savings Plans no Gerenciamento de Custos da AWS para gerar novas recomendações de Savings Plans a qualquer momento.

Depois de analisar a flexibilidade das instâncias, você pode confirmar seguindo as recomendações. Crie uma modelagem de custos analisando os custos de curto prazo da workload com possíveis opções de recursos diferentes, analisando os modelos de preço da AWS e alinhando-os aos requisitos empresariais para encontrar o custo total de propriedade e otimização dos custos de custos.

Etapas da implementação

Executar uma análise de desconto de compromisso: use o Cost Explorer na conta, analise as recomendações de instâncias reservadas e Savings Plans. Entenda as recomendações de Savings Plans e estime seus gastos mensais e as economias mensais. Examine as recomendações no nível da conta de gerenciamento, que são calculadas considerando o uso em todas as contas em sua organização da AWS que têm o compartilhamento de descontos de RI ou Savings Plans habilitado, com o intuito de ter o máximo de economia nas contas. Você pode verificar se implementou as recomendações corretas com os descontos e riscos necessários seguindo os laboratórios do Well-Architected.

Recursos

Documentos relacionados:

- Como a definição de preço da AWS funciona?
- Opções de compra de instância
- Visão geral dos Savings Plans
- Recomendações de Savings Plans
- Acesso a recomendações de instância reservada

- Conceitos básicos sobre a recomendação de Savings Plans
- Como os Savings Plans se aplicam ao uso da AWS
- Saving Plans com faturamento consolidado
- Ativação de instâncias reservadas compartilhadas e descontos de Savings Plans

Vídeos relacionados:

- Economize até 90% e execute cargas de trabalho de produção no local

Exemplos relacionados:

- AWS Well-Architected Lab: Pricing Models (Level 200)
- AWS Well-Architected Labs: Pricing Model Analysis (Level 200)
- O que devo considerar antes de comprar um Savings Plan?
- How can I use rolling Savings Plans to reduce commitment risk?
- Quando uso instâncias spot

CUSTOS 8. Como planejar as cobranças de transferência de dados?

Planeje e monitore 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 Executar a modelagem de transferência de dados
- COST08-BP02 Selecionar os componentes para otimizar o custo da transferência de dados
- COST08-BP03 Implantar serviços para reduzir custos de transferência de dados

COST08-BP01 Executar a 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 exposição a riscos se esta prática recomendada não for estabelecida: alto

Orientações para a implementação

Ao projetar uma solução na nuvem, as taxas de transferência de dados geralmente são negligenciadas devido ao hábito de projetar a arquitetura usando datacenters on-premises ou à falta de conhecimento. As taxas de transferência de dados na AWS são determinadas pela origem, pelo destino e pelo volume do tráfego. A consideração dessas taxas durante a fase de projeto pode resultar em redução de custos. É muito importante compreender onde ocorre a transferência de dados na workload, o custo da transferência e os respectivos benefícios associados para estimar com precisão o custo total de propriedade (TCO). 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 os componentes dos serviços que transferem os dados na workload e conclui que esse é um custo aceitável (de modo semelhante ao pagamento por computação e armazenamento nas duas 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.

Ao modelar a transferência de dados, considere a quantidade de dados ingeridos e a origem desses dados. Além disso, considere a quantidade de dados processados e a quantidade de armazenamento ou capacidade computacional necessária. Durante a modelagem, siga as práticas recomendadas de rede para sua arquitetura de workload a fim de otimizar os possíveis custos de transferência de dados.

O AWS Pricing Calculator pode ajudar você a ver os custos estimados de serviços específicos da AWS e da transferência de dados esperada. Se você já tiver uma workload em execução (para fins de teste ou em um ambiente de pré-produção), use o AWS Cost Explorer ou o AWS Cost and Usage Report (CUR) para compreender e modelar os 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

- Identificar os requisitos: qual é a principal meta e os requisitos de negócios para a transferência de dados planejada entre a origem e o destino? Quais são os resultados obtidos no final? Reúna os requisitos de negócios e defina o resultado esperado.
- Identificar a origem e o destino: qual é a fonte de dados e o destino da transferência de dados; por exemplo, dentro das Regiões da AWS, para os serviços da AWS ou para a internet?

  - Data transfer within an Região da AWS
  - Data transfer between Regiões da AWS
  - Data transfer out to the internet
- Identificar as classificações dos dados: qual é a classificação dos dados para essa transferência de dados? De que tipo são esses dados? Qual é o tamanho dos dados? Com que frequência os dados devem ser transferidos? Os dados são sigilosos?
- Identificar as ferramentas ou os serviços da AWS a serem usados: quais serviços da AWS são usados para essa transferência de dados? É possível usar um serviço já provisionado para outra workload?
- Calcular os custos da transferência de dados: use os Preços da AWS e a modelagem de transferência de dados que você criou anteriormente para calcular os custos da transferência de dados para a workload. Calcule os custos da transferência de dados em diferentes níveis de uso, tanto para aumentos quanto para reduções no uso da workload. Quando houver várias opções para a arquitetura da workload, calcule o custo de cada uma delas a título de comparação.
- Vincular os custos aos resultados: para cada custo de transferência de dados incorrido, especifique o resultado obtido pela workload. Se a transferência for entre componentes, poderá ser para desacoplamento; se for entre zonas de disponibilidade, poderá ser para redundância.
- Criar modelagem de transferência de dados: depois de coletar todas as informações, crie uma modelagem de transferência de dados de base conceitual para vários casos de uso e diferentes workloads.

Recursos

Documentos relacionados:

- Soluções de armazenamento em cache da AWS
- Preços da AWS
- Preços do Amazon EC2
- Preços da Amazon VPC
- Understanding data transfer charges

Vídeos relacionados:

- Monitoramento e otimização dos custos de transferência de dados

- Introduction to Amazon S3 Transfer Acceleration

Exemplos relacionados:

- Overview of Data Transfer Costs for Common Architectures (Visão geral dos custos de transferência de dados para arquiteturas comuns)
- AWS Prescriptive Guidance for Networking

COST08-BP02 Selecionar os componentes para otimizar o custo da 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 exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

A arquitetura da transferência de dados minimiza os custos da transferência de dados. Isso pode envolver o uso de redes de entrega de conteúdo para localizar os dados mais perto dos usuários ou o uso de links de rede dedicados do ambiente on-premises para a AWS. Você também pode usar a otimização de WAN e a otimização de aplicações para reduzir a quantidade de dados transferidos entre componentes.

Ao transferir dados para a Nuvem AWS ou dentro dela, é essencial conhecer o destino com base em diversos casos de uso, a natureza dos dados e os recursos de rede disponíveis para selecionar os serviços certos da AWS e otimizar a transferência de dados. A AWS oferece uma variedade de serviços de transferência de dados personalizados para diversos requisitos de migração de dados. Selecione as opções corretas de armazenamento de dados e de transferência de dados com base nas necessidades empresariais da organização.

Ao planejar ou analisar a arquitetura da workload, considere o seguinte:

- Usar os endpoints da VPC na AWS: os endpoints da VPC permitem conexões privadas entre a VPC e os serviços da AWS compatíveis. Isso permite evitar o uso da internet pública, o que pode resultar em custos de transferência de dados.
- Usar um gateway NAT: use um gateway NAT para que as instâncias em uma sub-rede privada possam se conectar à internet ou aos serviços fora da VPC. Verifique se os recursos por trás do

gateway NAT que enviam mais tráfego estão na mesma zona de disponibilidade do gateway NAT. Caso contrário, crie novos gateways NAT na mesma zona de disponibilidade do recurso para reduzir as taxas de transferência de dados entre AZs.

- Usar o AWS Direct Connect: o AWS Direct Connect ignora a internet pública e estabelece uma conexão direta e privada entre a rede on-premises e a AWS. Isso pode ser mais econômico e consistente do que transferir grandes volumes de dados pela internet.
- Evitar transferir dados entre limites regionais: as transferências de dados entre Regiões da AWS (de uma região para outra) normalmente geram cobranças. A decisão de seguir um caminho multirregional deve ser muito cuidadosa. Para obter mais detalhes, consulte Cenários de várias regiões.
- Monitorar a transferência de dados: use o Amazon CloudWatch e os logs de fluxo da VPC para capturar detalhes sobre a transferência de dados e o uso da rede. Analise as informações de tráfego de rede capturadas nas VPCs, como o endereço IP ou o intervalo de entrada e saída das interfaces de rede.
- Analisar o uso da rede: use ferramentas de medição e de geração de relatórios, como os painéis de CUDOS do AWS Cost Explorer ou o CloudWatch, para compreender o custo da transferência de dados da workload.

Etapas da implementação

- Selecionar os componentes da transferência de dados: usando a modelagem de transferência de dados explicada em COST08-BP01 Executar a modelagem de transferência de dados, concentre-se em quais são os maiores custos da transferência de dados ou em quais seriam se o uso da workload mudasse. Procure arquiteturas alternativas ou componentes adicionais que removam ou reduzam a necessidade da transferência de dados (ou que diminuam o custo).

Recursos

Práticas recomendadas relacionadas:

- COST08-BP01 Executar a modelagem de transferência de dados
- COST08-BP03 Implantar serviços para reduzir custos de transferência de dados

Documentos relacionados:

- Migração de dados para a nuvem

- Soluções de armazenamento em cache da AWS
- Deliver content faster with Amazon CloudFront

Exemplos relacionados:

- Overview of Data Transfer Costs for Common Architectures (Visão geral dos custos de transferência de dados para arquiteturas comuns)
- AWS Network Optimization Tips
- Optimize performance and reduce costs for network analytics with VPC Flow Logs in Apache Parquet format (Otimize o desempenho e reduza os custos de análise da rede com os logs de fluxo da VPC no formato Apache Parquet)

COST08-BP03 Implantar serviços para reduzir custos de transferência de dados

Implemente serviços para reduzir a transferência de dados. Por exemplo, é possível usar locais da borda ou redes de entrega de conteúdo (CDN) para fornecer conteúdo aos usuários finais, criar camadas de cache na frente de servidores de aplicações ou bancos de dados e usar conexões de rede dedicadas em vez de VPN para conectividade com a nuvem.

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

Orientação para implementação

Existem vários serviços da AWS que podem ajudar a otimizar o uso da transferência de dados pela rede. Dependendo da arquitetura da nuvem e dos componentes e tipo da workload, esses serviços podem ajudar na compactação, no armazenamento em cache e no compartilhamento e na distribuição do tráfego na nuvem.

- 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 um grande número de usuários globalmente com latência mínima. O pacote security savings pode ajudar você a economizar até 30% do uso do CloudFront se você planeja aumentar o uso ao longo do tempo.
- 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 a rede privada e a rede global da AWS. Ele é ideal para pequenos escritórios ou parceiros de negócios porque oferece conectividade simplificada, além de ser um serviço totalmente gerenciado e elástico.
- Endpoints da VPC permitem conectividade entre os serviços da AWS em redes privadas e podem ser usados para reduzir os custos de transferência de dados pública e Gateway NAT NAT. VPC endpoints de gateway não tem cobranças por hora e oferecem suporte ao 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.
- gateways fornecem escalabilidade e gerenciamento integrados, reduzindo os custos, em comparação com uma instância NAT independente. Coloque os gateways NAT nas mesmas zonas de disponibilidade das instâncias de alto tráfego e pense no uso de endpoints da VPC para as instâncias que precisam acessar o Amazon DynamoDB ou o Amazon S3 a fim de reduzir os custos de transferência e processamento de dados.
- Use AWS Snow Family dispositivos que têm recursos de computação para coletar e processar dados na borda. Os dispositivos da AWS Snow Family (Snowcone, o Snowball e Snowmobile) permitem que você mova petabytes de dados para o ambiente econômico e off-line da Nuvem AWS.

Etapas da implementação

- Implementar serviços: Selecione os serviços de rede aplicáveis da AWS com base no serviço e no tipo de workload usando a modelagem de transferência de dados e revisando os logs de fluxo da VPC. Veja onde estão os maiores custos e os maiores fluxos de volume. Analise os serviços da AWS e avalie se algum deles reduz ou remove a transferência, especificamente a entrega de conteúdo e as 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
- AWS Explore Our Products
- AWS caching solutions
- Amazon CloudFront
- AWS Snow Family

- Pacote Amazon CloudFront Security Savings

Vídeos relacionados:

- Monitoramento e otimização dos custos de transferência de dados
- Série de otimização de custos da AWS: CloudFront
- Como posso reduzir as taxas de transferência de dados do meu gateway NAT?

Exemplos relacionados:

- How-to chargeback shared services: An AWS Transit Gateway example
- Understand AWS data transfer details in depth from cost and usage report using Athena query and QuickSight
- Overview of Data Transfer Costs for Common Architectures (Visão geral dos custos de transferência de dados para arquiteturas comuns)
- Using AWS Cost Explorer to analyze data transfer costs
- Cost-Optimizing your AWS architectures by utilizing Amazon CloudFront features
- Como posso reduzir as taxas de transferência de dados do meu gateway NAT?

Gerenciar recursos de demanda e fornecimento

Pergunta

- CUSTOS 9. Como gerenciar a demanda e fornecer recursos?

CUSTOS 9. Como gerenciar a demanda e fornecer recursos?

Para uma workload 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 Realizar uma análise sobre a demanda da workload

- COST09-BP02 Implementar um buffer ou controle de utilização para gerenciar a demanda
- COST09-BP03 Fornecer recursos dinamicamente

COST09-BP01 Realizar uma análise sobre a demanda da workload

Analise a demanda da workload 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 trabalho 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 para implementação

Analisar a demanda de workload para computação em nuvem envolve entender os padrões e as características das tarefas de computação que são iniciadas no ambiente de nuvem. Essa análise ajuda os usuários a otimizar a alocação de recursos, gerenciar custos e verificar se a performance atende aos níveis exigidos.

Conhecer os requisitos da workload. Os requisitos da organização devem indicar os tempos de resposta da workload para solicitações. O tempo de resposta pode ser usado para determinar se a demanda é gerenciada ou se a oferta de recursos deve ser 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. Realize a análise durante um período longo o suficiente para incorporar qualquer variação sazonal, como processamento de fim de mês ou picos de fim de ano.

O trabalho de análise deve refletir os possíveis benefícios da implementação do ajuste de escala. Observe o custo total esperado do componente e os aumentos ou diminuições no uso e no custo durante a vida útil da workload.

Veja abaixo alguns aspectos importantes a serem considerados ao realizar a análise da demanda de workload para computação em nuvem:

1. Métricas de utilização e performance de recursos: analise como os recursos da AWS estão sendo usados ao longo do tempo. Determine padrões de uso de pico e fora do pico para otimizar as estratégias de alocação e ajuste de escala de recursos. Monitore métricas de performance, como tempos de resposta, latência, throughput e taxas de erro. Essas métricas ajudam a avaliar a integridade geral e a eficiência da infraestrutura de nuvem.

2. Comportamento de escalabilidade de usuários e aplicações: entenda o comportamento do usuário e como ele afeta a demanda da workload. Examinar os padrões de tráfego de usuários ajuda a aprimorar a entrega de conteúdo e a capacidade de resposta das aplicações. Analise como as workloads escalam com o aumento da demanda. Determine se os parâmetros de ajuste de escala automático estão configurados de forma correta e eficaz para lidar com flutuações de carga.
3. Tipos de workload: identifique os diferentes tipos de workload em execução na nuvem, como processamento em lote, processamento de dados em tempo real, aplicação web, bancos de dados ou machine learning. Cada tipo de workload pode ter requisitos de recursos e perfis de performance diferentes.
4. Acordos de serviço (SLAs): compare a performance real com os SLAs para garantir a conformidade e identificar áreas que precisam ser aprimoradas.

Você pode usar o Amazon CloudWatch para coletar e monitorar métricas, monitorar arquivos de log, definir alarmes e reagir automaticamente a mudanças nos recursos da AWS. Você também pode usar o Amazon CloudWatch para obter visibilidade sobre a utilização de recursos, a performance das aplicações e a integridade operacional em todo o sistema.

Com o AWS Trusted Advisor, é possível provisionar os recursos seguindo as práticas recomendadas para melhorar a performance e a confiabilidade do sistema, aumentar a segurança e procurar oportunidades de economia. Também é possível desativar o uso e as instâncias de não produção e usar o Amazon CloudWatch e o Auto Scaling para equiparar aumentos ou reduções na demanda.

Finalmente, você pode usar o AWS Cost Explorer ou Amazon QuickSight com o arquivo do AWS Cost and Usage Report (CUR) ou os logs da aplicação para realizar análises avançadas da demanda de workload.

No geral, uma análise abrangente da demanda da workload permite que as organizações tomem decisões embasadas sobre provisionamento, ajuste de escala e otimização de recursos, o que melhora a performance, o custo-benefício e a satisfação do usuário.

Etapas para a 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 oAmazon CloudWatch, os arquivos de log e os dados de monitoramento para obter informações sobre como a workload foi usada. Analise um ciclo completo da workload e colete dados para alterações sazonais, como eventos de fim de mês ou de ano. O esforço refletido na análise deve refletir as características da workload. Deve-se concentrar o maior esforço em workloads de alto valor

com as maiores alterações na demanda. Por outro lado, deve-se concentrar o menor esforço em workloads de baixo valor que tenham alterações mínimas na demanda.

- 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 possam alterar a demanda da workload. Preveja a demanda da workload com esses dados.

Recursos

Documentos relacionados:

- Amazon CloudWatch
- AWS Trusted Advisor
- AWS X-Ray
- AWS Auto Scaling
- O AWS Programador de Instâncias
- Conceitos básicos do Amazon SQS
- AWS Cost Explorer
- Amazon QuickSight

Vídeos relacionados:

Exemplos relacionados:

- Monitorar, acompanhar e analisar em prol da otimização de custos
- Searching and analyzing logs in CloudWatch

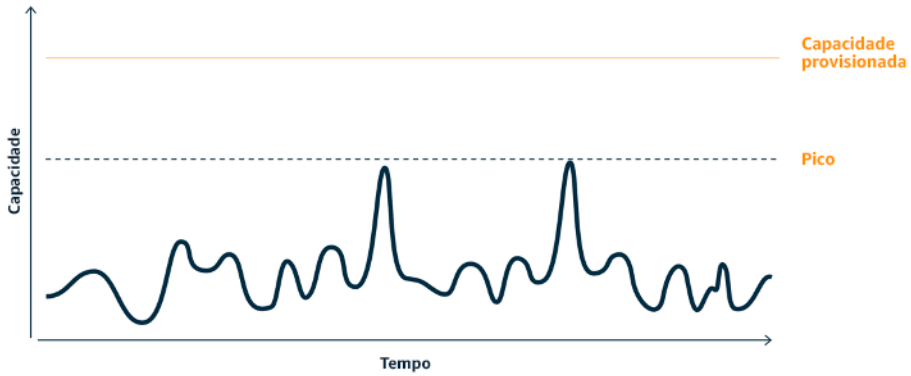
COST09-BP02 Implementar um buffer ou controle de utilização para gerenciar a demanda

O armazenamento em buffer e o controle de utilização modificam a demanda na carga de trabalho, suavizando todos os picos. Implemente o controle de utilização quando seus clientes realizarem novas tentativas. Implemente o armazenamento em buffer para armazenar a solicitação e adiar o processamento até um momento posterior. Verifique se os controles de utilização e buffers estão projetados para que os clientes recebam uma resposta no tempo necessário.

Nível de exposição a riscos se esta prática recomendada não for estabelecida: médio

Orientações para a implementação

A implementação de um buffer ou controle de utilização é crucial na computação em nuvem para gerenciar a demanda e reduzir a capacidade provisionada necessária para a workload. Para um desempenho ideal, é essencial avaliar a demanda total, incluindo os picos, o ritmo das mudanças nas solicitações e o tempo de resposta necessário. Quando os clientes têm a capacidade de reenviar solicitações, é prático aplicar o controle de utilização. Entretanto, para clientes que não têm funcionalidades de repetição, a abordagem ideal é implementar uma solução de buffer. Esses buffers agilizam o influxo de solicitações e otimizam a interação de aplicações com velocidades operacionais variadas.



Curva da demanda com dois picos distintos que exigem alta capacidade provisionada

Considere uma workload com a curva de demanda mostrada na figura anterior. Essa workload tem dois picos e, para lidar com eles, é provisionada a capacidade de recurso mostrada pela linha laranja. Os recursos e a energia usados para essa workload não são indicados pela área abaixo da curva da demanda, mas pela área abaixo da linha da capacidade provisionada, visto que é preciso ter capacidade provisionada para lidar com esses dois picos. Nivelar a curva da demanda pode ajudar você a reduzir a capacidade provisionada para uma workload e a diminuir o respectivo impacto ambiental. Para suavizar o pico, considere implementar uma solução de controle de utilização ou de buffer.

Para entendê-los melhor, vamos examinar o controle de utilização e o buffer.

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 não for possível atender à solicitação no momento, ela deverá tentar novamente mais tarde. A origem espera por um período e repete 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, é possível usar o Amazon API Gateway para implementar o controle de utilização.

Baseado em buffer: uma abordagem baseada em buffer usa produtores (componentes que enviam mensagens para a fila), consumidores (componentes que recebem mensagens da fila) e uma fila (que contém mensagens) para armazenar as mensagens. As mensagens são lidas pelos consumidores e processadas, permitindo que as mensagens sejam executadas na taxa que atenda aos requisitos de negócios dos consumidores. Usando uma metodologia centrada em buffer, as mensagens dos produtores são armazenadas em filas ou fluxos, prontas para serem acessadas pelos consumidores em um ritmo alinhado às demandas operacionais.

Na AWS, é possível escolher entre vários serviços para implementar uma abordagem de buffer. O Amazon Simple Queue Service (Amazon SQS) é um serviço gerenciado que fornece filas que permitem que um único consumidor leia mensagens individuais. O Amazon Kinesis fornece um fluxo que permite que muitos consumidores leiam as mesmas mensagens.

O buffer e o controle de utilização podem suavizar qualquer pico modificando a demanda da workload. Use o controle de utilização quando os clientes repetirem ações e use o buffer para reter a solicitação e processá-la posteriormente. Ao trabalhar com uma arquitetura com uma abordagem baseada em buffer, arquitete a workload para atender à solicitação no tempo necessário e verifique se é possível lidar com solicitações duplicadas de trabalho. Analise a demanda geral, a taxa de alteração e o tempo de resposta necessário para dimensionar adequadamente o controle ou buffer necessário.

Etapas da implementação

- Analisar os requisitos do cliente: analise as solicitações de cliente para determinar se eles podem realizar novas tentativas. Para clientes que não podem realizar novas tentativas, será necessário implementar buffers. 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 um controle de utilização: implemente um buffer ou um controle de utilização na workload. Uma fila, como o Amazon Simple Queue Service (Amazon SQS), pode fornecer um buffer para os componentes da workload. O Amazon API Gateway pode fornecer o controle de utilização para os componentes da workload.

Recursos

Práticas recomendadas relacionadas:

- SUS02-BP06 Implementar armazenamento em buffer ou controle de utilização para nivelar a curva da demanda
- REL05-BP02 Controlar a utilização de solicitações

Documentos relacionados:

- AWS Auto Scaling
- AWS Instance Scheduler
- Amazon API Gateway
- Amazon Simple Queue Service
- Conceitos básicos do Amazon SQS
- Amazon Kinesis

Vídeos relacionados:

- Escolha do serviço de mensagem correto para sua aplicação distribuída

Exemplos relacionados:

- Managing and monitoring API throttling in your workloads (Gerenciar e monitorar o controle de utilização de API em workloads)
- Throttling a tiered, multi-tenant REST API at scale using API Gateway
- Enabling Tiering and Throttling in a Multi-Tenant Amazon EKS SaaS Solution Using Amazon API Gateway
- Integração de aplicações usando filas e mensagens

COST09-BP03 Fornecer recursos dinamicamente

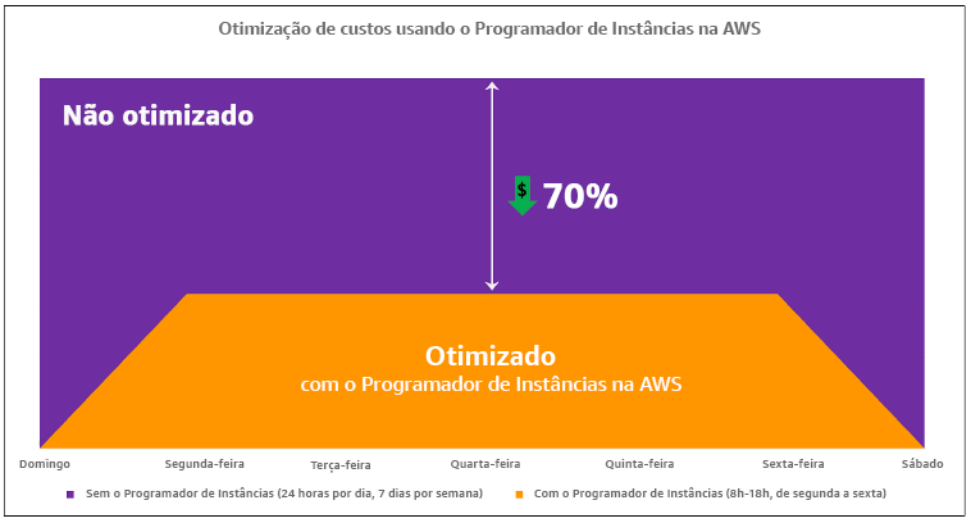
Os recursos são provisionados de maneira planejada. Isso pode ser baseado na demanda, como por meio da escalabilidade automática, ou no tempo, em que a demanda é previsível e os recursos são fornecidos com base no tempo. Esses métodos ocasionam a menor quantidade de superprovisionamento ou subprovisionamento.

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

Orientação para implementação

Há várias maneiras de os clientes da AWS aumentarem os recursos disponíveis para suas aplicações e fornecerem recursos para atender à demanda. Uma dessas opções é usar o AWS Programador de Instâncias, que automatiza a inicialização e a interrupção das instâncias do Amazon Elastic Compute Cloud (Amazon EC2) e do Amazon Relational Database Service (Amazon RDS). A outra opção é usar o AWS Auto Scaling, que possibilita escalar automaticamente seus recursos de computação com base na demanda de sua aplicação ou serviço. O fornecimento de recursos com base na demanda permitirá que você pague somente pelos recursos utilizados, reduza os custos lançando recursos quando eles forem necessários e os encerre quando não forem.

O AWS Programador de Instâncias possibilita configurar a interrupção e o início de suas instâncias do Amazon EC2 e do Amazon RDS em horários definidos para que você possa atender à demanda pelos mesmos recursos em um padrão de tempo consistente, por exemplo, acesso diário dos usuários às instâncias do Amazon EC2 às 8h, que não são necessárias após as 18h. Essa solução ajuda a reduzir o custo operacional interrompendo recursos que não estão sendo usados e iniciá-los quando eles são necessários.



Otimização de custos com o AWS Programador de Instâncias.

Você também pode configurar facilmente programações para suas instâncias do Amazon EC2 em suas contas e regiões com uma interface de usuário (IU) simples usando a Configuração rápida do AWS Systems Manager. É possível programar instâncias do Amazon EC2 e do Amazon RDS com o AWS Programador de Instâncias e interromper e iniciar instâncias existentes. No entanto, você não pode interromper e iniciar instâncias que façam parte de seu grupo do Auto Scaling (ASG) nem que

gerenciem serviços, como o Amazon Redshift ou o Amazon OpenSearch Service. Os grupos do Auto Scaling têm seu próprio agendamento para as instâncias no grupo e essas instâncias são criadas.

O AWS Auto Scaling ajuda você a ajustar sua capacidade para manter uma performance estável e previsível pelo menor custo possível para atender às variações de demanda. Trata-se de um serviço totalmente gerenciado e gratuito para escalar a capacidade de sua aplicação que se integra a instâncias do Amazon EC2 e frotas spot, Amazon ECS, Amazon DynamoDB e 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 ajuste de escala incorporadas para otimizar a performance, os custos ou um equilíbrio entre os dois, além de oferecer escalabilidade preditiva para ajudar com picos que ocorrem regularmente.

Há várias opções de ajuste de escala disponíveis para escalar seu grupo do Auto Scaling:

- Manter os níveis de instância atuais em todos os momentos
- Escalar manualmente
- Escalar com base em um cronograma
- Escalar com base na demanda
- Usar o ajuste de escala preditivo

As políticas do Auto Scaling diferem e podem ser categorizadas como políticas de ajuste de escala dinâmico e programado. As políticas dinâmicas são ajuste de escala manual ou dinâmico, ajuste de escala programado ou preditivo. Você pode usar políticas para ajuste de escala dinâmico, programado e preditivo. Também é possível usar métricas e alarmes do Amazon CloudWatch para acionar eventos de escalabilidade para sua workload. Recomendamos que você use modelos de lançamento, que permitem acessar os recursos e melhorias mais recentes. Nem todos os recursos do Auto Scaling estão disponíveis quando você usa as configurações de inicialização. Por exemplo, você não pode criar um grupo do Auto Scaling que inicie instâncias spot e sob demanda nem que especifique vários tipos de instância. Você deve usar um modelo de inicialização para configurar esses recursos. Ao usar modelos de inicialização, recomendamos que você crie a versão de cada um. Com o versionamento dos modelos de inicialização, você pode criar um subconjunto do conjunto completo de parâmetros. Depois, é possível reutilizá-lo para criar outras versões do mesmo modelo de inicialização.

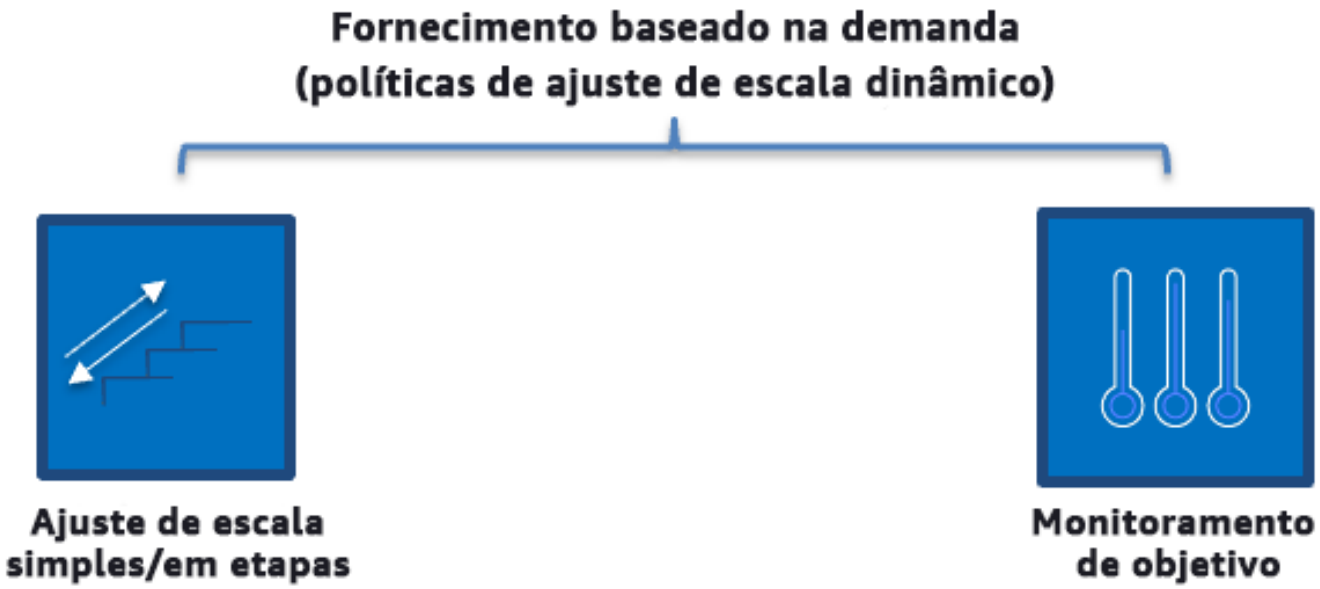
É possível usar o AWS Auto Scaling ou incorporar ajuste de escala em seu código com as APIs da AWS ou SDKs. Isso reduz os custos gerais da workload removendo o custo operacional de fazer alterações manualmente em seu ambiente, e alterações podem ser realizadas muito mais

rapidamente. Isso também atende à mobilização de recursos da workload de acordo com sua demanda a qualquer momento. Para seguir essa prática recomendada e fornecer recursos de forma dinâmica para sua organização, você precisa entender a escalabilidade horizontal e vertical na Nuvem AWS, bem como a natureza das aplicações executadas em instâncias do Amazon EC2. É melhor para sua equipe de gerenciamento financeiro na nuvem trabalhar com equipes técnicas a fim de seguir essa prática recomendada.

O Elastic Load Balancing (Elastic Load Balancing) ajuda a escalar distribuindo a demanda entre vários recursos. Com o uso do ASG e do Elastic Load Balancing, você pode gerenciar as solicitações recebidas roteando o tráfego de forma ideal para que nenhuma instância fique sobrecarregada em um grupo do Auto Scaling. As solicitações seriam distribuídas entre todos os destinos de um grupo-alvo de forma contínua, sem considerar a capacidade nem a utilização.

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. Quando possível, você deve usar uma métrica que seja indicativa da experiência do cliente. Normalmente é uma métrica personalizada que pode se originar do código da aplicação em sua workload. Para elaborar como atender à demanda dinamicamente neste documento, vamos agrupar o Auto Scaling em duas categorias, como modelos de fornecimento baseados na demanda e baseados no tempo, e nos aprofundarmos em cada uma delas.

Fornecimento baseado em demanda: Utilize a elasticidade da nuvem para fornecer recursos para atender às mudanças na demanda, confiando no estado de demanda quase em tempo real. Para fornecimento baseado em demanda, use as APIs ou os recursos de serviço para variar programaticamente a quantidade de recursos de nuvem em sua arquitetura. Isso permite que você ajuste a escala de componentes em sua arquitetura e aumente o número de recursos durante picos de demanda a fim de manter a performance e reduzir a capacidade quando a demanda diminui para reduzir os custos.

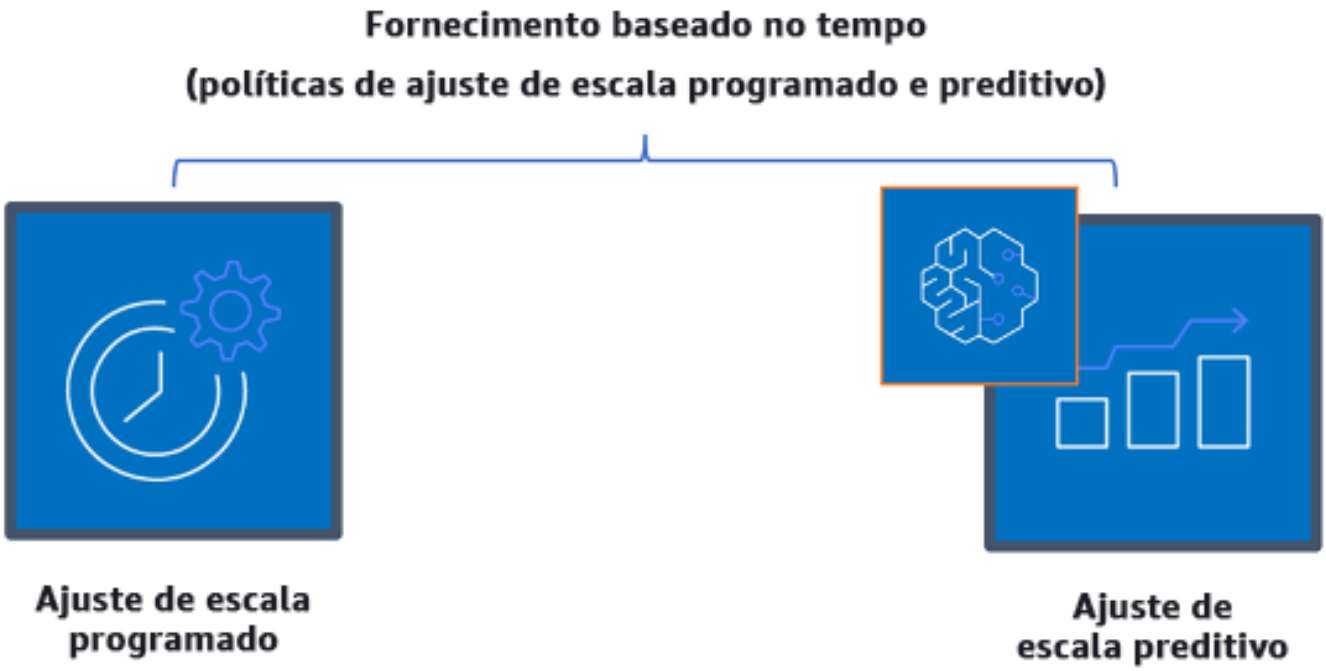


Políticas de ajuste de escala dinâmico com base na demanda

- Ajuste de escala simples/em etapas: Monitora métricas e adiciona/remove instâncias de acordo com as etapas definidas manualmente pelos clientes.
- Monitoramento de objetivo: Mecanismo de controle semelhante a um termostato que adiciona ou remove instâncias automaticamente para manter as métricas em uma meta definida pelo cliente.

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.

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.



Políticas de ajuste de escala baseado em tempo

Você pode usar o ajuste de escala automático programado ou preditivo para implementar uma abordagem baseada em tempo. As workloads podem ser programadas para aumentar ou reduzir a escala horizontalmente em horários definidos (por exemplo, o início do horário comercial), tornando os recursos disponíveis quando os usuários chegarem ou a demanda aumentar. A escalabilidade preditiva usa padrões para aumentar a escala horizontalmente enquanto a escalabilidade programada usa horários predefinidos para isso. Você também pode usar a estratégia de seleção de tipo de instância baseada em atributos (ABS) em grupos do Auto Scaling, o que permite que você expresse seus requisitos de instância como um conjunto de atributos, como vCPU, memória e armazenamento. Isso permite usar automaticamente os tipos de instância de geração mais recente quando eles são lançados e acessar uma variedade mais ampla de capacidade com instâncias spot do Amazon EC2. A frota do Amazon EC2 e o Amazon EC2 Auto Scaling selecionam e executam instâncias que se ajustam aos atributos especificados, eliminando a necessidade de escolher manualmente os tipos de instância.

Você também pode aproveitar as APIs da AWS e os SDKs e o 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 workload 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 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 workloads 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 o ajuste de escala programado: Para alterações previsíveis na demanda, o ajuste de escala baseado 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. Para configurar a programação baseada em tempo, você pode usar o ajuste de escala preditivo do ajuste de escala programado para aumentar o número de instâncias do Amazon EC2 em seus grupos do Auto Scaling com antecedência de acordo com as alterações de carga esperadas ou previsíveis.
- Configure o ajuste de escala preditivo: O ajuste de escala preditivo permite aumentar com antecedência o número de instâncias do Amazon EC2 em seu grupo do Auto Scaling de padrões diários e semanais nos fluxos de tráfego. Se você tiver picos de tráfego regulares e aplicações que levem muito tempo para serem iniciadas, considere usar a escalabilidade preditiva. A escalabilidade preditiva pode ajudar você a escalar com maior rapidez inicializando a capacidade antes da carga projetada em comparação com a escalabilidade dinâmica isolada, que é reativa por natureza. Por exemplo, se os usuários começarem a usar sua workload no início do horário comercial e não usá-la após o expediente, a escalabilidade preditiva poderá adicionar capacidade antes do horário comercial, o que elimina o atraso da escalabilidade dinâmica para reagir à mudança no tráfego.
- Configure o ajuste de escala automático dinâmico: Para configurar o ajuste de escala com base nas métricas ativas da workload, use o Auto Scaling. Use a análise e configure o Auto Scaling para iniciar nos níveis de recursos corretos e garanta que a workload escale no tempo necessário. Você pode iniciar e dimensionar automaticamente uma frota de instâncias sob demanda e instâncias

spot em um único grupo do Auto Scaling. Além de receber descontos pelo uso de instâncias spot, você pode usar instâncias reservadas ou um Savings Plan para receber taxas com desconto do preço regular da instância sob demanda. Todos esses fatores combinados ajudam você a otimizar sua economia de custos para instâncias do Amazon EC2 e determinar a escala e a performance desejadas para sua aplicação.

Recursos

Documentos relacionados:

- AWS Auto Scaling
- O AWS Programador de Instâncias
- Escalar o tamanho de seu grupo do Auto Scaling
- Conceitos básicos do Amazon EC2 Auto Scaling
- Conceitos básicos do Amazon SQS
- Ajuste de escala programado para Amazon EC2 Auto Scaling
- Ajuste de escala preditivo para Amazon EC2 Auto Scaling

Vídeos relacionados:

- Políticas de ajuste de escala com monitoramento de objetivo para o Auto Scaling
- O AWS Programador de Instâncias

Exemplos relacionados:

- Attribute based Instance Type Selection for Auto Scaling for Amazon EC2 Fleet (Seleção de tipo de instância baseada em atributo do Auto Scaling para a frota do Amazon EC2)
- Optimizing Amazon Elastic Container Service for cost using scheduled scaling (Otimização do Amazon Elastic Container Service para o custo usando ajuste de escala programado)
- Ajuste de escala com o Amazon EC2 Auto Scaling
- How do I use Instance Scheduler with AWS CloudFormation to schedule Amazon EC2 instances? (Como usar o Programador de Instâncias com o AWS CloudFormation para programar as instâncias do Amazon EC2?)

Otimizar ao longo do tempo

Perguntas

- CUSTOS 10. Como avaliar os novos serviços?
- CUSTOS 11. Como avaliar o custo do esforço?

CUSTOS 10. Como avaliar 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 Revisar 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 workload. 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 ou a cada seis meses, enquanto workloads abaixo de 10% são analisadas anualmente.

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

Orientações para a implementação

Para ter a workload mais econômica, você deve revisar regularmente a workload para saber se há oportunidades de implementar novos serviços, recursos e componentes. Para obter custos gerais mais baixos, o processo deve ser proporcional à quantidade potencial de economia. Por exemplo, as workloads que representam 50% do seu gasto geral devem ser analisadas com mais frequência e mais precisão do que as workloads que representam 5% do seu gasto geral. Leve em consideração quaisquer fatores externos ou volatilidade. Se a workload 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 workload 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 workload, você estaria movendo um sistema ainda maior para a nova linguagem, exigindo ainda mais esforço do que anteriormente.

Divida sua workload 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 workload. 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.

Defina um processo para avaliar novos serviços, padrões de design, tipos de recursos e configurações para otimizar o custo de sua workload conforme ficarem disponíveis. Semelhante aos processos de análise de pilar de performance e análise de pilar de confiabilidade, identifique, valide e priorize as atividades de otimização e aprimoramento e correção de problemas e incorpore isso ao seu backlog.

Etapas da implementação

- Definição da frequência de análise: defina a frequência com que a workload e os componentes dela devem ser analisados. Aloque tempo e recursos para o aprimoramento contínuo e analise a frequência para melhorar a eficiência e a otimização de sua workload. Essa é uma combinação de fatores e pode diferir de workload para workload em sua organização e entre componentes na workload. 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 geraria 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 workload e os componentes dela. Você pode decidir analisar a workload completa a cada 18 meses, analisar os servidores web a cada seis meses, o banco de dados a cada doze meses, a

computação e o armazenamento de curto prazo a cada seis meses e o armazenamento de longo prazo a cada doze meses.

- Definição da minuciosidade da análise: defina quanto esforço é gasto na análise da workload ou dos componentes dela. Semelhante à frequência da análise, esse é um equilíbrio de vários fatores. Avalie e priorize oportunidades de melhorias para concentrar os esforços nos locais onde eles oferecem os maiores benefícios enquanto calcula quanto esforço é necessário para essas atividades. Se os resultados esperados não satisfizerem às metas e o esforço necessário custar mais, itere usando cursos de ação alternativos. Seus processos de análise devem incluir tempo e recursos dedicados para possibilitar melhorias incrementais contínuas. Por exemplo, você pode decidir gastar uma semana de análise no componente do banco de dados, uma semana de análise para recursos computacionais 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

Exemplos relacionados:

- AWS Support Proactive Services (Serviços proativos do AWS Support)
- Regular workload reviews for SAP workloads (Análises regulares de workloads SAP)

COST10-BP02 Revisar e analisar a workload regularmente

As workloads existentes são revisadas regularmente com base em cada processo definido para descobrir se é possível adotar novos serviços, substituir serviços já em vigor ou reprojetar workloads.

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

Orientações para a implementação

A AWS sempre adiciona novos recursos para que você possa experimentar e novar mais rápido com a tecnologia mais recente. Novidades da AWS detalha como a AWS está fazendo isso e oferece uma breve visão geral dos anúncios de expansão regional, dos recursos e dos serviços da AWS assim que eles são lançados. Você pode examinar detalhadamente os lançamentos que foram

anunciados e usá-los para revisar e analisar suas workloads existentes. Para obter os benefícios de novos serviços e recursos da AWS, analise suas workloads e implemente novos serviços e recursos conforme necessário. Isso significa que você pode precisar substituir os serviços que você usa para a workload ou modernizar a workload para adotar novos serviços da AWS. Por exemplo, você pode analisar suas workloads e substituir o componente de mensagens pelo Amazon Simple Email Service. 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.

Para analisar sua workload e destacar possíveis oportunidades, você deve considerar não apenas novos serviços, mas também novas formas de criar soluções. Examine os vídeos Esta é a minha arquitetura na AWS para conhecer designs de arquitetura de outros clientes, seus desafios e suas soluções. Confira a série Tudo incluído para descobrir aplicações reais dos serviços da AWS e conhecer histórias de clientes. Você também pode assistir à série de vídeo De volta ao básico, que explica, examina e detalha práticas recomendadas do padrão de arquitetura de nuvem básica. Outra fonte são os vídeos Como construir isso, que são projetados para ajudar as pessoas em grandes ideias sobre como viabilizar seu produto mínimo viável (MVP) usando serviços da AWS. Desse modo, criadores do mundo inteiro que tiverem uma grande ideia poderão obter orientações arquiteturais de arquitetos de soluções experientes da AWS. Por fim, você pode examinar os materiais do recurso Conceitos básicos, que tem tutoriais detalhados.

Antes de executar seu processo de avaliação, siga os requisitos de sua empresa com relação a workload, segurança e privacidade dos dados para usar requisitos específicos de serviço ou de região e performance e, ao mesmo tempo, siga o processo de avaliação que foi acordado.

Etapas da implementação

- Avaliação regular da workload: usando o processo definido, realize avaliações na frequência especificada. Verifique se você despendeu 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, e não apenas os benefícios de longo prazo.
- Implementação de novos serviços: se o resultado da análise for implementar alterações, primeiro execute uma linha de base da workload para saber o custo atual por saída. Implemente as alterações e, em seguida, execute uma análise para confirmar o novo custo por saída.

Recursos

Documentos relacionados:

- Blog de novidades da AWS
- Quais as novidades da AWS
- Documentação da AWS
- Conceitos básicos da AWS
- Recursos gerais da AWS

Vídeos relacionados:

- AWS: Esta é a minha arquitetura
- AWS: De volta ao básico
- AWS: Série Tudo Incluído
- Como construir isso

CUSTOS 11. Como avaliar o custo do esforço?

Práticas recomendadas

- COST11-BP01 Realizar automações nas operações

COST11-BP01 Realizar automações nas operações

Avalie o custo de esforço das operações na nuvem. Redução da quantidade de tempo e esforço em tarefas administrativas, implantação e outras operações usando a automação. Avalie o tempo e custo necessários ao esforço de operações e à automatização de tarefas administrativas para reduzir o esforço humano quando possível.

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

A automatização de processos melhora a consistência e a escalabilidade, oferece maior visibilidade, confiabilidade e flexibilidade, reduz os custos e acelera a inovação ao liberar recursos humanos e aperfeiçoar as métricas. Ela reduz a frequência das tarefas manuais, melhora a eficiência e beneficia as empresas por fornecer uma experiência consistente e confiável em implantação, administração ou operação de workloads. Você pode liberar recursos de infraestrutura de tarefas operacionais manuais e usá-los para tarefas e inovações de maior valor e, consequentemente, melhorar os resultados dos negócios. As empresas necessitam de um método comprovado e testado para gerenciar suas workloads na nuvem. Essa solução precisa ser segura, rápida e econômica, e oferecer risco mínimo e confiabilidade máxima.

Primeiro, priorize suas operações com base no esforço necessário examinando o custo geral das operações na nuvem. Por exemplo, quanto tempo se leva para implantar novos recursos na nuvem, realizar alterações de otimização nos recursos existentes ou implementar as configurações necessárias? Examine o custo total das ações humanas incluindo o custo de operações e gerenciamento como fator. Priorize a automação das tarefas administrativas para reduzir o esforço humano. A avaliação do esforço deve refletir o provável benefício. Por exemplo, tempo gasto na execução de tarefas manuais em contraposição a tarefas automáticas. Priorize a automatização de atividades de alto valor repetitivas. As atividades que apresentam um risco maior de erro humano normalmente são o melhor lugar para começar a automatizar porque, com frequência, o risco cria um custo operacional adicional não desejado (como horas extras de trabalho da equipe de operações).

Usando as ferramentas e os serviços da AWS ou produtos de terceiros, você pode escolher quais automações da AWS deve implementar e personalizar para suas necessidades específicas. A tabela abaixo mostra alguns recursos e funções essenciais de operações que você pode obter com os serviços da AWS para automatizar a administração e operação:

- AWS Audit Manager: audite continuamente seu uso da AWS para simplificar a avaliação de risco e conformidade
- AWS Backup: gerencie centralmente e automatize a proteção de dados.
- AWS Config: configure recursos de computação, avalie, audite e estime o valor das configurações e do inventário de recursos.
- AWS CloudFormation: lance recursos altamente disponíveis com infraestrutura como código.
- AWS CloudTrail: gerenciamento de mudanças de TI, conformidade e controle.
- Amazon EventBridge: programe eventos e acione o AWS Lambda para que ele tome medidas.
- AWS Lambda: automatize processos repetitivos acionando-os com eventos ou executando-os em uma programação fixa com o Amazon EventBridge.
- AWS Systems Manager: inicie e interrompa workloads, aplique patches em sistemas operacionais e automatize a configuração e o gerenciamento contínuo.
- AWS Step Functions: programe trabalhos e automatize fluxos de trabalho.
- AWS Service Catalog: crie modelos de consumo e infraestrutura como código com conformidade e controle.

Considere a economia de tempo que permitirá que sua equipe se concentre na retirada de recursos de endividamento técnico, inovação e agregação de valor. Por exemplo, talvez você precise mover sem alterações (lift-and-shift) seu ambiente on-premises para a nuvem o mais rapidamente possível

e otimizar em outro momento. Vale a pena explorar as economias que você poderia obter usando serviços totalmente gerenciados da AWS que eliminam ou reduzem custos de licença, como Amazon Relational Database Service, Amazon EMR, Amazon WorkSpaces e Amazon SageMaker. 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.

Se você quiser adotar automações imediatamente usando produtos e serviços da AWS e se não tiver habilidades em sua organização, entre em contato com o AWS Managed Services (AMS), AWS Professional Services ou com parceiros da AWS para ampliar a adoção da automação e melhorar sua excelência operacional na nuvem.

AWS Managed Services (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.

O AWS Professional Services também pode ajudar você a alcançar os resultados de negócios que deseja e a automatizar as operações com a AWS. Com o AWS Professional Services, você tem acesso a práticas de especialidade globais para apoiar suas iniciativas em áreas focalizadas de computação em nuvem empresarial. As disciplinas especializadas oferecem orientações direcionadas por meio de práticas recomendadas, frameworks, ferramentas e serviços em áreas do conhecimento industrial, de solução e de tecnologia. Elas ajudam os cliente a implantar operações de TI automatizadas, robustas e ágeis, bem como recursos de governança otimizados para o centro da nuvem.

Etapas da implementação

- Construir uma vez e implantar várias: use infraestrutura como código como o AWS CloudFormation, AWS SDKs ou a AWS Command Line Interface (AWS CLI) para implantar uma vez e usar várias vezes para o mesmo ambiente ou para cenários de recuperação de desastres. Marque e, ao mesmo tempo, monitore seu consumo tal como definido em outras práticas recomendadas. Use o AWS Launch Wizard para reduzir o tempo para implantar várias workloads empresariais conhecidas. O AWS Launch Wizard orienta você sobre tamanho, configuração e implantação de workloads empresariais seguindo práticas recomendadas da AWS. Também é possível usar o AWS Service Catalog, que ajuda você a criar e gerenciar modelos aprovados de infraestrutura como código para uso na AWS. Desse modo, qualquer pessoa pode descobrir recursos de nuvem de autoatendimento aprovados.

- Automatizar operações: execute operações de rotina automaticamente sem intervenção humana. Usando as ferramentas e os serviços da AWS, você pode escolher quais automações da AWS deve implementar e personalizar para suas necessidades específicas. Por exemplo, use EC2 Image Builder para criar, testar e implantar imagens de máquina virtual e contêiner para uso na AWS ou em ambiente on-premises. Se a ação que você deseja não puder ser realizada com serviços da AWS ou você precisar de ações mais complexas com recursos de filtragem, automatize suas operações usando a AWS CLI ou ferramentas dos AWS SDKs. A AWS CLI oferece a possibilidade de automatizar o processo completo de controle e gerenciamento de serviços da AWS por meio de scripts sem usar o Console da AWS. Selecione os AWS SDKs de sua preferência para interagir com os serviços da AWS. Para outros exemplos de código, consulte Repositório de exemplos de código de AWS SDKs.

Recursos

Documentos relacionados:

- Modernização de operações na Nuvem AWS
- Serviços da AWS para automação
- AWS Systems Manager Automation
- Automações da AWS para administração e operações SAP
- AWS Managed Services
- AWS Professional Services
- Infraestrutura e automação

Exemplos relacionados:

- Reinvenção das operações automatizadas (Parte I)
- Reinvenção das operações automatizadas (Parte II)
- Automações da AWS para administração e operações SAP
- Automações de TI com o AWS Lambda
- Repositório de exemplos de código da AWS
- Amostras da AWS

Sustentabilidade

O pilar Sustentabilidade abrange ações como compreender os impactos dos serviços usados, quantificá-los ao longo do ciclo de vida da workload e aplicar princípios e práticas recomendadas de design para reduzi-los ao criar workloads na nuvem. Você pode encontrar orientações prescritivas sobre implementação no Whitepaper sobre o pilar Sustentabilidade.

Áreas de práticas recomendadas

- Seleção de região
- Alinhamento com a demanda
- Software e arquitetura
- Dados
- Hardware e serviços
- Processo e cultura

Seleção de região

Pergunta

- SUS 1 Como selecionar regiões para sua workload?

SUS 1 Como selecionar regiões para sua workload?

A escolha da região para sua workload afeta significativamente seus KPIs, incluindo desempenho, custo e pegada de carbono. Para melhorar efetivamente esses KPIs, você deve escolher regiões para suas workloads com base em requisitos empresariais e metas de sustentabilidade.

Práticas recomendadas

- SUS01-BP01 Escolher a região com base nos requisitos empresariais e nas metas de sustentabilidade

SUS01-BP01 Escolher a região com base nos requisitos empresariais e nas metas de sustentabilidade

Escolha uma região para sua workload com base em seus requisitos empresariais e metas de sustentabilidade para otimizar seus KPIs, incluindo desempenho, custo e pegada de carbono.

Antipadrões comuns:

- Selecione a região da workload com base em sua localização.
- Você consolida todos os recursos da workload em uma única localização geográfica.

Benefícios de estabelecer esta prática recomendada: Colocar uma workload próxima a projetos de energia renovável da Amazon ou regiões com baixa intensidade de carbono publicada pode ajudar a reduzir a pegada de carbono de uma workload na nuvem.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientações para a implementação

O Nuvem AWS é uma rede em constante expansão de regiões e pontos de presença (PoP), com uma infraestrutura de rede global que os conecta. A escolha da região para sua workload afeta significativamente seus KPIs, incluindo desempenho, custo e pegada de carbono. Para melhorar efetivamente esses KPIs, você deve escolher regiões para sua workload com base em seus requisitos empresariais e metas de sustentabilidade.

Etapas da implementação

- Siga estas etapas para avaliar e selecionar possíveis regiões para sua workload com base em seus requisitos de negócios, incluindo conformidade, recursos disponíveis, custo e latência:
  - Confirme se essas regiões estão em conformidade, com base nos regulamentos locais exigidos.
  - Use as Listas de serviços regionais do AWS para verificar se as regiões têm os serviços e recursos necessários para executar sua workload.
  - Calcule o custo da workload em cada região usando o AWS Pricing Calculator.
  - Teste a latência de rede entre as localizações de seus usuários finais e cada Região da AWS.
- 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).
  - Identifique suas diretrizes de sustentabilidade relevantes para rastrear e comparar as emissões de carbono ano a ano com base no GHG Protocol (métodos baseados no mercado e baseados na localização).
  - Escolha a região com base no método que você usa para rastrear as emissões de carbono. Para obter mais detalhes sobre como escolher uma região com base em suas diretrizes de sustentabilidade, consulte Como selecionar uma região para sua workload com base nas metas de sustentabilidade.

Recursos

Documentos relacionados:

- Compreensão de suas estimativas de emissão de carbono
- Amazon em todo o mundo
- Metodologia de energia renovável
- O que considerar ao selecionar uma região para suas workloads

Vídeos relacionados:

- Arquitetura sustentável e redução de sua pegada de carbono do AWS

Alinhamento com a demanda

Pergunta

- SUS 2 Como alinhar recursos de nuvem à sua demanda?

SUS 2 Como alinhar recursos de nuvem à sua demanda?

A maneira como os usuários e as aplicações consomem suas workloads e outros recursos pode ajudar você a identificar melhorias para atingir as metas de sustentabilidade. Escale a infraestrutura de forma que ela corresponda à demanda e use apenas os recursos mínimos necessários para oferecer suporte aos usuários. Alinhe os níveis de serviço às necessidades do cliente. Posicione os recursos a fim de limitar a rede necessária para que usuários e aplicações os consumam. Elimine ativos não utilizados. Forneça aos membros da sua equipe dispositivos compatíveis com suas necessidades e minimize o impacto na sustentabilidade.

Práticas recomendadas

- SUS02-BP01 Escalar a infraestrutura da workload dinamicamente
- SUS02-BP02 Alinhar os SLAs com as metas de sustentabilidade
- SUS02-BP03 Interromper a criação e a manutenção de ativos não utilizados
- SUS02-BP04 Otimizar o posicionamento geográfico das workloads com base nos respectivos requisitos de rede
- SUS02-BP05 Otimizar os recursos dos membros da equipe para as atividades realizadas

- SUS02-BP06 Implementar armazenamento em buffer ou controle de utilização para nivelar a curva da demanda

SUS02-BP01 Escalar a infraestrutura da workload dinamicamente

Use a elasticidade da nuvem e escale sua infraestrutura de forma dinâmica para corresponder a oferta de recursos de nuvem à demanda e evitar capacidade superprovisionada em sua workload.

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 dessa prática recomendada: configurar e testar a elasticidade da workload ajuda a corresponder de maneira eficiente a oferta de recursos de nuvem à demanda e evitar a capacidade superprovisionada. Você pode aproveitar a elasticidade na nuvem para escalar automaticamente a capacidade durante e depois de picos de demanda para garantir que esteja usando apenas o número exato de recursos necessários para atender aos requisitos do seu negócio.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientações para a implementação

A nuvem fornece a flexibilidade de expandir ou reduzir seus recursos dinamicamente por meio de diversos mecanismos para atender a mudanças na demanda. O equilíbrio ideal entre a oferta e a demanda oferece o menor impacto ambiental para uma workload.

A demanda pode ser fixa ou variável, exigindo métricas e automação para garantir que o gerenciamento não se torne um gasto excessivo. Os aplicativos podem aumentar ou diminuir a escala verticalmente ao modificar o tamanho da instância e horizontalmente ao modificar o número de instâncias, ou uma combinação de ambos.

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

- Abordagem de monitoramento de meta: monitore sua métrica de escalabilidade e aumente ou diminua automaticamente a capacidade conforme necessário.

- Escalabilidade preditiva: escale antecipadamente em relação às tendências diárias e semanais.
- Abordagem com base na programação: defina sua própria programação de escalabilidade de acordo com as alterações de carga previsíveis.
- Escalabilidade de serviços: escolha serviços (como tecnologia sem servidor) que são escalados nativamente por design ou fornecem escalabilidade automática como um recurso.

Identifique períodos de utilização baixa ou sem utilização e escale os recursos para eliminar a capacidade em excesso e melhorar a eficiência.

Etapas da implementação

- A elasticidade corresponde à oferta de recursos que você tem face à demanda por estes recursos. Instâncias, contêineres e funções fornecem mecanismos para elasticidade, seja em combinação com a escalabilidade automática ou como um recurso do serviço. A AWS fornece uma variedade de mecanismos de escalabilidade automática para garantir que as workloads possam reduzir a escala verticalmente de forma rápida e fácil durante períodos de baixa carga de usuário. Veja alguns exemplos de mecanismos de escalabilidade automática:

Auto scaling mechanism	Where to use
Amazon EC2 Auto Scaling	Use para verificar se você tem o número correto de instâncias do Amazon EC2 disponíveis para processar a carga de usuário para o seu aplicativo.
Application Auto Scaling	Use para escalar 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).
o dimensionador automático de cluster do Kubernetes	Use para escalar automaticamente os clusters do Kubernetes na AWS.

- A escalabilidade geralmente é discutida em relação a serviços de computação, como instâncias do Amazon EC2 ou funções do AWS Lambda. Considere a configuração de serviços não

relacionados a computação, como o Amazon DynamoDB, e grave unidades de capacidade ou fragmentos do Amazon Kinesis Data Streams para corresponder à demanda.

- 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.
- Use a escalabilidade dinâmica em vez da escalabilidade manual para o seu grupo do Auto Scaling. Também recomendamos que você use as políticas de escalabilidade de monitoramento de meta na sua escalabilidade dinâmica.
- Verifique se as implantações da workload podem lidar com eventos de aumento e redução horizontal da escala. Crie cenários de teste para eventos de redução horizontal da escala para verificar se a workload se comporta conforme o esperado e não afeta a experiência do usuário (como perda da sessão persistente). Você também pode usar o histórico de atividades para verificar a 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. Com a escalabilidade preditiva, é possível eliminar a necessidade de superprovisionar a capacidade. Para obter mais detalhes, consulte Escalabilidade preditiva com o Amazon EC2 Auto Scaling.

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?
- Monitorar a carga do banco de dados com o Performance Insights no Amazon RDS

- Introdução ao suporte nativo para escalabilidade preditiva com o Amazon EC2 Auto Scaling
- Apresentando o Karpenter: um dimensionador automático de clusters do Kubernetes de código aberto e alta performance
- Aprofundamento do Amazon ECS Cluster Auto Scaling

Vídeos relacionados:

- Build a cost-, energy-, and resource-efficient compute environment (Criar um ambiente de computação eficiente em termos de custo, energia e recursos)
- 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

Analise e otimize os Acordos de Serviço (SLA) com base em suas metas de sustentabilidade para minimizar os recursos necessários a fim de oferecer compatibilidade com sua workload enquanto continua a atender às necessidades empresariais.

Antipadrões comuns:

- SLAs de workload são desconhecidos ou ambíguos.
- Você define seu SLA apenas para disponibilidade e performance.
- Você usa o mesmo padrão de design (como arquitetura multi-AZ) para todas as suas workloads.

Benefícios do estabelecimento desta prática recomendada: o alinhamento dos SLAs com as metas de sustentabilidade ocasiona o uso ideal dos recursos e a concretização das necessidades empresariais.

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

Orientações para a implementação

Os SLAs definem o nível de serviço esperado de uma workload de nuvem, como tempo de resposta, disponibilidade e retenção de dados. Eles influenciam a arquitetura, o uso de recursos e o impacto ambiental de uma workload de nuvem. Em uma cadência regular, analise os SLAs e faça compensações que reduzam significativamente o uso de recursos em troca de reduções aceitáveis em níveis de serviço.

Etapas da implementação

- Defina ou remodele SLAs que apoiem suas metas de sustentabilidade e, ao mesmo tempo, atendam aos seus requisitos empresariais, não os excedendo.
- Faça compensações que reduzam significativamente os impactos na sustentabilidade em troca de reduções aceitáveis em níveis de serviço.
  - Sustentabilidade e confiabilidade: workloads altamente disponíveis tendem a consumir mais recursos.
  - Sustentabilidade e performance: o uso de mais recursos para impulsionar a performance pode causar um maior impacto ambiental.
  - Sustentabilidade e segurança: workloads excessivamente seguras podem ter um maior impacto ambiental.
- Use padrões de design, como microsserviços na AWS 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 da AWS (SLAs)
- Importance of Service Level Agreement for SaaS Providers

Vídeos relacionados:

- Delivering sustainable, high-performing architectures (Entregar arquiteturas sustentáveis e de alta performance)
- Criar um ambiente de computação eficiente em termos de custo, energia e recursos

SUS02-BP03 Interromper a criação e a manutenção de ativos não utilizados

Desative ativos em sua workload para reduzir o número de recursos necessários para atender à sua demanda e minimizar o desperdício.

Antipadrões comuns:

- Você não analisa sua aplicação com relação a ativos redundantes ou não mais necessários.
- Você não remove ativos redundantes ou não mais necessários.

Benefícios do estabelecimento desta prática recomendada: a remoção de ativos ociosos libera recursos e melhora a eficiência geral da workload.

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

Orientações para a implementação

Os ativos ociosos consomem recursos de nuvem como espaço de armazenamento e potência computacional. Com a identificação e eliminação desses ativos, você pode liberar esses recursos e aumentar a eficiência da arquitetura de nuvem. Analise regularmente 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. Remova esses ativos redundantes para diminuir o desperdício de recursos em sua workload.

Etapas da implementação

- Use ferramentas de monitoramento para identificar ativos estáticos que não são mais necessários.
- Antes de remover qualquer ativo, avalie o impacto da remoção sobre a arquitetura.
- Desenvolva um plano e remova os ativos que não são mais necessários.
- Consolide ativos gerados sobrepostos para remover o processamento redundante.
- Atualize suas aplicações para que não produzam nem armazenem mais ativos que não são necessários.
- 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.
- Avalie regularmente sua workload para identificar e remover ativos ociosos.

Recursos

Documentos relacionados:

- Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte II: Armazenamento
- Como faço para verificar se há recursos ativos dos quais não preciso mais na minha Conta da AWS?

Vídeos relacionados:

- Como verifico e depois removo recursos ativos dos quais não preciso mais na minha Conta da AWS?

SUS02-BP04 Otimizar o posicionamento geográfico das workloads com base nos respectivos requisitos de rede

Selecione locais e serviços de nuvem para sua workload que reduzam a distância que o tráfego de rede deve percorrer e diminua o total de recursos de rede necessários para comportar a workload.

Antipadrões comuns:

- Selecione a região da workload com base em sua localização.
- Você consolida todos os recursos da workload em uma única localização geográfica.
- Todo o tráfego flui por meio dos datacenters existentes.

Benefícios de estabelecer esta 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 para implementação

A infraestrutura da Nuvem AWS é construída em torno de opções de local, como regiões, zonas de disponibilidade, grupos de posicionamento e locais da borda, como AWS Outposts e zonas locais da AWS. Essas opções de local são responsáveis por manter a conectividade entre componentes da aplicação, serviços de nuvem, redes da borda e datacenters on-premises.

Analise os padrões de acesso à rede em sua workload para identificar como usar essas opções de local de nuvem e reduzir a distância que o tráfego de rede precisa percorrer.

Etapas da implementação

- Analise os padrões de acesso à rede em sua workload para identificar como os usuários utilizam sua aplicação.
  - Use ferramentas de monitoramento, como Amazon CloudWatch e o AWS CloudTrail, para coletar dados sobre as atividades da rede.
  - Analise os dados para identificar o padrão de acesso à rede.
- Selecione as regiões para implantação da workload com base nos seguintes elementos fundamentais:
  - 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 (ou regiões) próxima dos clientes de sua workload.
  - Outras restrições: Leve em conta restrições, como custo e conformidade, conforme explicado em O que considerar ao selecionar uma região para suas workloads.
- Use armazenamento em cache local ou soluções de armazenamento em cache da AWS para ativos usados com frequência a fim de aumentar a performance, reduzir a movimentação de dados e reduzir o impacto ambiental.

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

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

Service	Quando usar
o Lambda@Edge	Use para operações com uso intenso de computação que são iniciadas quando objetos não estão no cache.
Funções do Amazon CloudFront	Use para casos de uso simples, como solicitações HTTP(s) ou manipulações de resposta, que podem ser iniciadas por funções de curta duração.
AWS IoT Greengrass	Use para executar computação local, mensagens e armazenamento de dados em cache para dispositivos conectados.

- 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

Vídeos relacionados:

- Demystifying data transfer on AWS (Desmistificação da transferência de dados na AWS)
- Escalar a performance da rede em instâncias do Amazon EC2 de última geração

Exemplos relacionados:

- Workshops de redes da AWS
- Arquitetura para a sustentabilidade: reduza a movimentação de dados entre redes

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 ambiental e, ao mesmo tempo, atender às suas necessidades.

Antipadrões comuns:

- Você ignora o impacto dos dispositivos usados pelos membros da equipe sobre a eficiência geral de sua aplicação de nuvem.
- Você gerencia e atualiza manualmente os recursos usados pelos membros da equipe.

Benefícios do estabelecimento desta prática recomendada: otimizar recursos para os membros da equipe melhora a eficiência geral das aplicações habilitadas para a nuvem.

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

Orientações para a implementação

Conheça os recursos que os membros da equipe usam para consumir seus serviços, o ciclo de vida esperado e o impacto financeiro e na sustentabilidade. Implemente estratégias para otimizar esses recursos. Por exemplo, realize operações complexas, como renderização e compilação, em infraestrutura escalável com alta utilização em vez de em sistemas de usuário único subutilizados com alto consumo de energia.

Etapas da implementação

- Provisione estações de trabalho e outros dispositivos para alinhar a maneira como eles são usados.
- Use áreas de trabalho 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 a fim de utilizar a respectiva elasticidade.

- 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.
  - O AWS Systems Manager Fleet Manager é uma experiência de interface do usuário (UI) que ajuda você a gerenciar remotamente os nós em execução na AWS ou no ambiente on-premises.

Recursos

Documentos relacionados:

- O que é o Amazon WorkSpaces?
- Otimizador de custos para o Amazon WorkSpaces
- Documentação do Amazon AppStream 2.0
- NICE DCV

Vídeos relacionados:

- Gerenciamento de custos do Amazon WorkSpaces na AWS

SUS02-BP06 Implementar armazenamento em buffer ou controle de utilização para nivelar a curva da demanda

O armazenamento em buffer e o controle de utilização nivelam a curva da demanda e reduzem a capacidade provisionada necessária para sua workload.

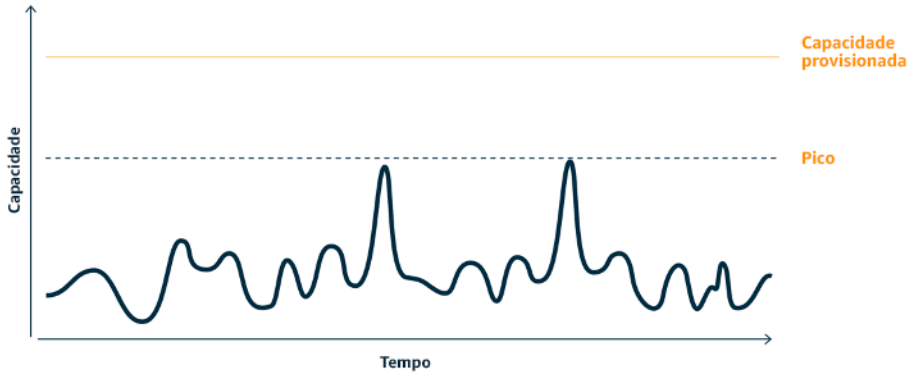
Antipadrões comuns:

- Você processa imediatamente as solicitações de cliente embora isso não seja necessário.
- Você não analisa os requisitos das solicitações de cliente.

Benefícios do estabelecimento desta prática recomendada: nivelar a curva da demanda reduz a capacidade provisionada necessária para a workload. Reduzir a capacidade provisionada significa diminuir o consumo de energia e o impacto ambiental.

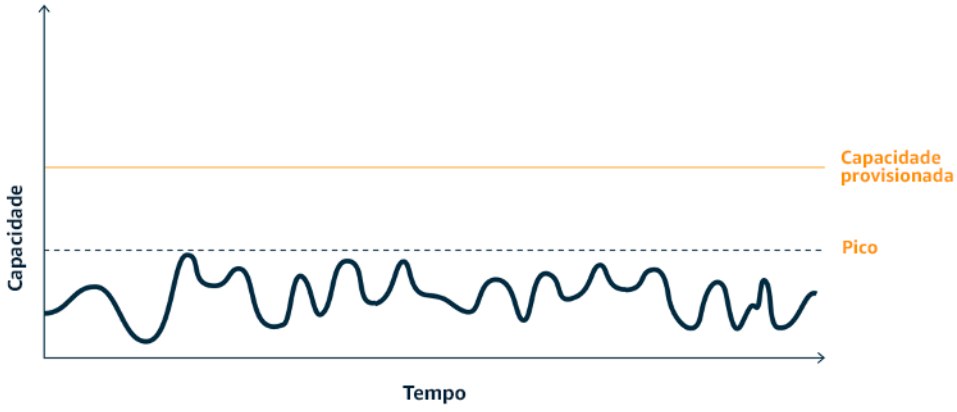
Nível de exposição a riscos quando esta prática recomendada não é estabelecida: baixo

Nivelar a curva da demanda pode ajudar você a reduzir a capacidade provisionada para uma workload e a diminuir o respectivo impacto ambiental. Considere a workload com a curva de demanda mostrada na figura a seguir. Essa workload tem dois picos e, para lidar com eles, é provisionada a capacidade de recurso mostrada pela linha laranja. Os recursos e energia usados para essa workload não são indicados pela área abaixo da curva da demanda, mas pela área abaixo da linha da capacidade provisionada, visto que é preciso ter capacidade provisionada para lidar com esses dois picos.



Curva da demanda com dois picos distintos que exigem alta capacidade provisionada.

Você pode usar o armazenamento em buffer ou o controle de utilização para modificar a curva da demanda e atenuar os picos, o que significa menor capacidade provisionada e menor consumo de energia. Implemente o controle de utilização quando seus clientes puderem realizar novas tentativas. Implemente o armazenamento em buffer para armazenar a solicitação e adiar o processamento até um momento posterior.



Efeito do controle de utilização sobre a curva da demanda e a capacidade provisionada.

Etapas da implementação

- Analise as solicitações dos clientes para determinar como responder a elas. As perguntas a serem consideradas incluem:
  - Essa solicitação pode ser processada assincronamente?
  - O cliente tem capacidade de repetição?
- Se o cliente tiver capacidade de repetição, você pode implementar o controle de utilização, que informa à origem que, se ela não puder atender à solicitação naquele momento, deverá tentar novamente mais tarde.
  - Você pode usar o Amazon API Gateway para implementar o controle de utilização.
- Para clientes que não podem realizar novas tentativas, é necessário implementar um buffer para nivelar a curva da demanda. O buffer adia o processamento de solicitações, permitindo que as aplicações executadas em diferentes taxas se comuniquem com eficácia. Uma abordagem baseada em buffer usa uma fila ou um fluxo para aceitar mensagens 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.
  - O Amazon Simple Queue Service (Amazon SQS) é um serviço gerenciado que fornece filas que permitem que um único consumidor leia mensagens individuais.
  - O Amazon Kinesis oferece um fluxo que permite que vários consumidores leiam as mesmas mensagens.
- Analise a demanda geral, a taxa de alteração e o tempo de resposta necessário para dimensionar adequadamente o controle ou buffer necessário.

Recursos

Documentos relacionados:

- Conceitos básicos do Amazon SQS
- Integração de aplicações usando filas e mensagens

Vídeos relacionados:

- Escolha do serviço de mensagem correto para sua aplicação distribuída

Software e arquitetura

Pergunta

- SUS 3 Como aproveitar os padrões de software e arquitetura para apoiar as metas de sustentabilidade?

SUS 3 Como aproveitar os padrões de software e arquitetura para apoiar as 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
- SUS03-BP02 Remover ou refatorar componentes da workload subutilizados ou não utilizados
- SUS03-BP03 Otimizar as áreas de código que consomem mais tempo ou recursos
- SUS03-BP04 Otimizar o impacto sobre dispositivos e equipamentos
- SUS03-BP05 Usar arquiteturas e padrões de software que atendam melhor aos padrões de armazenamento e acesso aos dados

SUS03-BP01 Otimizar o software e a arquitetura para trabalhos assíncronos e programados

Use software eficiente e padrões de arquitetura, como orientado a filas, para manter uma alta e consistente utilização dos recursos implantados.

Antipadrões comuns:

- Provisione em excesso os recursos em sua workload na nuvem para atender a picos imprevistos na demanda.

- Sua arquitetura não separa remetentes e destinatários de mensagens assíncronas por um componente de sistema de mensagens.

Benefícios do estabelecimento desta prática recomendada:

- Padrões eficientes de software e arquitetura minimizam os recursos não utilizados em sua workload e melhoram a eficiência geral.
- Você pode dimensionar o processamento independentemente do recebimento de mensagens assíncronas.
- Por meio de um componente de mensagens, você relaxou os requisitos de disponibilidade que podem ser atendidos com menos recursos.

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

Orientações para a implementação

Use padrões de arquitetura eficientes, como arquitetura orientada por eventos, que resultam na utilização uniforme de componentes e minimizam o provisionamento em excesso em sua workload. A utilização de padrões de arquitetura eficientes minimiza recursos ociosos por falta de uso devido a mudanças na demanda ao longo do tempo.

Entenda os requisitos de seus componentes de workload e adote padrões de arquitetura que aumentam a utilização geral dos recursos. Retire os componentes que não são mais necessários.

Etapas da implementação

- Analise a demanda de sua workload para determinar como responder a ela.
- Para solicitações ou trabalhos que não exigem respostas síncronas, use arquiteturas orientadas por filas e operadores de escalonamento automático para maximizar a utilização. Aqui estão alguns exemplos de quando você pode considerar a arquitetura orientada por filas:

Queuing mechanism	Description
Filas de trabalho do AWS Batch	Os trabalhos do AWS Batch são enviados para uma fila de trabalhos onde permanece m até que possam ser agendados para execução em um ambiente de computação.

Queuing mechanism	Description
Instâncias spot do Amazon Simple Queue Service e Amazon EC2	Emparelhamento de Amazon SQS e instância s spot para criar uma arquitetura eficiente e tolerante a falhas.

- Para solicitações ou trabalhos que podem ser processados a qualquer momento, use mecanismos de agendamento para processar trabalhos em lote para maior eficiência. Aqui estão alguns exemplos de mecanismos de agendamento no AWS:

Scheduling mechanism	Description
Agendador do Amazon EventBridge	A capacidade do Amazon EventBridge que permite criar, executar e gerenciar tarefas agendadas em escala.
Programação baseada em tempo do AWS Glue	Defina uma programação baseada em tempo para seus crawlers e trabalhos no AWS Glue.
Tarefas agendadas do Amazon Elastic Container Service (Amazon ECS)	O Amazon ECS permite a criação de tarefas agendadas. Tarefas agendadas usam regras do Amazon EventBridge para executar tarefas em um agendamento ou em resposta a um evento do EventBridge.
Programador de Instâncias	Configure programações de início e parada para suas instâncias do Amazon EC2 e Amazon Relational Database Service.

- Se você usar mecanismos de pesquisa e webhooks em sua arquitetura, substitua-os por eventos. Use arquiteturas orientadas por eventos para criar workloads altamente eficientes.
- Aproveite a a computação sem servidor no AWS para eliminar a infraestrutura provisionada em excesso.
- Dimensione corretamente componentes individuais de sua arquitetura para evitar recursos ociosos aguardando entrada.

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?
- Usando o AWS Lambda com o Amazon SQS
- O que é o Amazon EventBridge?

Vídeos relacionados:

- Moving to event-driven architectures (Mudando para arquiteturas orientadas a eventos)

SUS03-BP02 Remover ou refatorar componentes da workload subutilizados ou não utilizados

Remova os componentes que não são mais utilizados nem necessários e refatore os componentes pouco usados para minimizar o desperdício em sua workload.

Antipadrões comuns:

- Você não verifica regularmente o nível de utilização de componentes individuais de sua workload.
- Você não verifica nem analisa as recomendações de ferramentas de dimensionamento correto da AWS, como o AWS Compute Optimizer.

Benefícios do estabelecimento desta prática recomendada: a remoção de ativos ociosos minimiza o desperdício e melhorar a eficiência geral da workload de nuvem.

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

Orientações para a implementação

Avalie sua workload para identificar componentes ociosos ou não utilizados. Esse é um processo de melhoria interativo que pode ser acionado por alterações na demanda ou pelo lançamento de um novo serviço de nuvem. Por exemplo, uma queda significativa no tempo de execução da função do

AWS Lambda pode ser uma indicação de que é necessário reduzir o tamanho da memória. Além disso, à medida que a AWS lança novos serviços e recursos, a arquitetura e os serviços ideais para sua workload podem mudar.

Monitore continuamente a atividade da workload e procure oportunidades para melhorar o nível de utilização de componentes individuais. Com a remoção de componentes ociosos e a execução de atividades de dimensionamento correto, você atende aos seus requisitos empresariais com menos recursos de nuvem.

Etapas da implementação

- Monitore e capture métricas de utiliza de componentes essenciais de sua workload (como utilização de CPU, utilização de memória ou throughput de rede nas métricas do Amazon CloudWatch).
- Para workloads estáveis, confira regularmente as ferramentas de dimensionamento correto da AWS, como o AWS Compute Optimizer, para identificar componentes ociosos, não usados ou subutilizados.
- Para workloads efêmeras, avalie as métricas de utilização para identificar componentes ociosos, não usados ou subutilizados.
- Retire componentes e ativos associados (como imagens do Amazon ECR) que não são mais necessários.
- Refatore ou consolide os componentes subutilizados com outros recursos para melhorar a eficiência da utilização. Por exemplo, você pode provisionar vários bancos de dados pequenos em uma única instância de banco de dados do Amazon RDS, em vez de executar bancos de dados em instâncias individuais subutilizadas.
- Saiba quais recursos são provisionados por sua workload para concluir uma unidade de trabalho.

Recursos

Documentos relacionados:

- AWS Trusted Advisor
- O que é o Amazon CloudWatch?
- Limpeza automatizada de imagens não utilizadas no Amazon ECR

Exemplos relacionados:

- Laboratório do Well-Architected: Dimensionamento correto com o AWS Compute Optimizer
- Laboratório do Well-Architected: Otimizar padrões de hardware e observar KPIs de sustentabilidade

SUS03-BP03 Otimizar as áreas de código que consomem mais tempo ou recursos

Otimize o código que é executado em diferentes componentes de sua arquitetura para minimizar o uso de recursos e, ao mesmo tempo, maximizar a performance.

Antipadrões comuns:

- Você ignora a otimização de seu código para uso de recursos.
- Normalmente, você responde a problemas de performance aumentando os recursos.
- Seu processo de revisão e desenvolvimento de código não monitora alterações na performance.

Benefícios de estabelecer esta prática recomendada: O uso de código eficiente minimiza o uso de recursos e melhora a performance.

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

Orientação para implementação

É essencial examinar toda área funcional, incluindo o código referente a uma aplicação projetada para a nuvem, para otimizar o uso de recursos e a performance. Monitore continuamente a performance da workload em ambientes de compilação e na produção e identifique oportunidades para melhorar os trechos cujo uso de recursos é particularmente alto. Adote um processo de revisão regular para identificar erros ou antipadrões dentro do código que usa os recursos ineficazmente. Utilize algoritmos simples e eficientes que produzem os mesmos resultados para seu caso de uso.

Etapas da implementação

- Ao desenvolver suas workloads, adote um processo de revisão de código automatizada para melhorar a qualidade e identificar erros e antipadrões.
  - Análises de código automatizadas com o Amazon CodeGuru Reviewer
  - Detecção de erros simultâneos com o Amazon CodeGuru
  - Elevação da qualidade do código para aplicações Python usando o Amazon CodeGuru
- À medida que você executa suas workloads, monitore os recursos para identificar componentes com altos requisitos de recurso por unidade de trabalho como alvos para revisões de código.

- Para revisões de código, 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.
  - Reduzir a pegada de carbono de sua organização com o Amazon CodeGuru Profiler
  - Conceitos básicos sobre o uso de memória em sua aplicação Java com o Amazon CodeGuru Profiler
  - Melhorar a experiência do cliente e reduzir os custos com o Amazon CodeGuru Profiler
- Use a linguagem de programação e o sistema operacional mais eficientes para a workload. Para obter detalhes sobre linguagens de programação com eficiência energética (incluindo Rust), consulte Sustentabilidade com Rust.
- Substitua os algoritmos com uso intenso de computação por uma versão mais simples e mais eficiente que produza o mesmo resultado.
- Remova códigos desnecessários, como classificações e formatações.

Recursos

Documentos relacionados:

- O que é o Amazon CodeGuru Profiler?
- Instâncias de FPGA
- Os AWS SDKs em Ferramentas para desenvolver na AWS

Vídeos relacionados:

- Improve Code Efficiency Using Amazon CodeGuru Profiler (Como melhorar a eficiência do código usando o Amazon CodeGuru Profiler)
- Automate Code Reviews and Application Performance Recommendations with Amazon CodeGuru (Como automatizar revisões de código e recomendações de performance de aplicação com o Amazon CodeGuru)

SUS03-BP04 Otimizar o impacto sobre dispositivos e equipamentos

Conheça os dispositivos e equipamentos usados em sua arquitetura e use estratégias para reduzir o respectivo uso. Isso pode minimizar o impacto ambiental de modo geral de sua workload de nuvem.

Antipadrões comuns:

- Você ignora o impacto ambiental dos dispositivos usados por seus clientes.
- Você gerencia e atualiza manualmente os recursos usados pelos clientes.

Benefícios do estabelecimento desta prática recomendada: implementar padrões e recursos de software que são otimizados para o dispositivo do cliente pode reduzir o impacto ambiental de modo geral da workload de nuvem.

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

Orientações para a implementação

Implementar padrões e recursos de software que são otimizados para os dispositivos do clientes pode reduzir o impacto ambiental de variadas maneiras:

- Implementar novos recursos que são compatíveis com versões anteriores pode reduzir o número de substituições de hardware.
- Otimizar uma aplicação para ser executada com eficiência nos dispositivos pode ajudar a reduzir o consumo de energia e a estender a duração da bateria (se eles forem alimentados por bateria).
- Otimizar uma aplicação para dispositivos também pode reduzir a transferência de dados ao longo da rede.

Conheça os dispositivos e equipamentos usados em sua arquitetura, o ciclo de vida esperado e o impacto da substituição desses componentes. Implemente padrões e recursos de software que possam ajudar a minimizar o consumo de energia do dispositivo, bem como a necessidade de os clientes substituírem o dispositivo e também atualizá-lo manualmente.

Etapas da implementação

- Faça um inventário dos dispositivos usados em sua arquitetura. Os dispositivos podem ser celular, tablet, dispositivos IoT, lâmpada inteligente ou até dispositivos inteligentes em uma fábrica.
- Otimize a aplicação executada nos dispositivos:
  - Use estratégias como execução de tarefas em segundo plano para reduzir o consumo de energia.
  - 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.
  - Converta cargas úteis e arquivos nos formatos otimizados exigidos pelos dispositivos. Por exemplo, você pode usar o Amazon Elastic Transcoder ou o AWS Elemental MediaConvert para

converter arquivos de mídia digital grandes e de alta qualidade em formatos que os usuários possam reproduzir em dispositivos móveis, tablets, navegadores Web e televisores conectados.

  - 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.
- Use um mecanismo sem fio automatizado para implantar atualizações em um ou mais dispositivos.
  - Você pode usar um pipeline de CI/CD para atualizar aplicativos móveis.
  - Você pode usar o AWS IoT Device Management para gerenciar remotamente dispositivos conectados em escala.
- Para testar novos recursos e atualizações, use parques de dispositivos gerenciados com conjuntos representativos de hardware e itere o desenvolvimento para maximizar os dispositivos compatíveis. Para obter mais detalhes, consulte SUS06-BP04 Usar parques de dispositivos gerenciados para testes.

Recursos

Documentos relacionados:

- O que é o AWS Device Farm?
- Documentação do Amazon AppStream 2.0
- NICE DCV
- Tutorial de OTA para atualização de firmware em dispositivos que executam o FreeRTOS

Vídeos relacionados:

- Introdução ao AWS Device Farm

SUS03-BP05 Usar arquiteturas e padrões de software que atendam melhor aos padrões de armazenamento e acesso aos dados

Entenda como os dados são usados com sua workload, consumidos pelos usuários, transferidos e armazenados. Use os padrões e arquiteturas de software ideais para acesso e armazenamento

de dados a fim de minimizar os recursos de computação, rede e armazenamento necessários para atender à workload.

Antipadrões comuns:

- Você pressupõe que todas as workloads têm padrões de acesso e armazenamento de dados semelhantes.
- Você usa apenas um nível de armazenamento, supondo que todas as workloads se encaixem nesse nível.
- Você pressupõe que os padrões de acesso aos dados permanecerão consistentes ao longo do tempo.
- Na eventualidade de uma alta expansão no acesso aos dados, sua arquitetura é capaz de comportá-la, mas isso faz com que os recursos fiquem ociosos na maior parte do tempo.

Benefícios do estabelecimento desta prática recomendada: selecionar e otimizar sua arquitetura com base nos padrões de acesso e armazenamento de dados ajudará a diminuir a complexidade do desenvolvimento e a aumentar a utilização de modo geral. Compreender quando usar tabelas globais, provisionamento de dados e armazenamento em cache ajuda a reduzir a despesas operacionais indiretas e a escalar com base nas necessidades da workload.

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

Orientações para a implementação

Use padrões de software e arquitetura que melhor se alinhem às características dos dados e aos padrões de acesso. Por exemplo, use uma arquitetura de dados moderna na AWS que permita que você use serviços com propósito específico otimizados para seus casos de uso exclusivos de análise. Esses padrões de arquitetura possibilitam um processamento de dados eficiente e reduzem o uso de recursos.

Etapas da implementação

- Analise as características dos dados e os padrões de acesso para identificar a configuração correta para seus recursos de nuvem. Principais características a serem consideradas:
  - Tipo de dados: estruturados, semiestruturados e não estruturados
  - Crescimento dos dados: delimitado, não delimitado
  - Durabilidade dos dados: persistentes, efêmeros, transitórios
  - Padrões de acesso: leituras ou gravações, frequência de atualização, com picos ou consistente

- Use padrões de arquitetura que comportem melhor os padrões de armazenamento e acesso aos dados.
  - Vamos arquitetar! Arquiteturas de dados modernas
  - Bancos de dados na AWS: a ferramenta certa para o trabalho certo
- Use tecnologias que funcionam nativamente com dados compactados.
- Use serviços de análise com propósito específico para processamento de dados em sua arquitetura.
- Use o mecanismo de banco de dados que melhor comporta seu padrão de consulta dominante. Gerencie seus índices de bancos de dados para garantir a execução eficiente de consultas. Para ter mais detalhes, consulte Bancos de dados da AWS.
- Escolha protocolos de rede que reduzam a quantidade de capacidade de rede consumida em sua arquitetura.

Recursos

Documentos relacionados:

- Suporte a compactação no Athena
- COPY de formatos de dados colunar com o Amazon Redshift
- Converter o formato de registro de entrada no Firehose
- Opções de formato de dados para entradas e saídas no AWS Glue
- Melhorar a performance de consultas no Amazon Athena com a conversão em formatos colunares
- Carregar arquivos de dados compactados do Amazon S3 com o Amazon Redshift
- Monitorar a carga de banco de dados com o Performance Insights no Amazon Aurora
- Monitorar a carga de banco de dados com o Performance Insights no Amazon RDS
- Classe de armazenamento do Amazon S3 Intelligent-Tiering

Vídeos relacionados:

- Criar arquiteturas modernas de dados na AWS

Dados

Pergunta

SUS 4 Como aproveitar as políticas e os padrões de gerenciamento de dados para apoiar as 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 promovam o valor empresarial dos dados de forma mais eficaz 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
- SUS04-BP02 Usar tecnologias compatíveis com seus padrões de acesso e de armazenamento de dados
- SUS04-BP03 Usar políticas para gerenciar o ciclo de vida de seus conjuntos de dados
- SUS04-BP04 Usar elasticidade e automação para expandir o armazenamento em bloco ou o sistema de arquivos
- SUS04-BP05 Remover dados desnecessários ou redundantes
- SUS04-BP06 Usar sistemas de arquivos compartilhados ou armazenamento para acessar dados comuns
- SUS04-BP07 Minimizar a movimentação de dados entre redes
- SUS04-BP08 Fazer backup de dados somente quando for difícil recriar

SUS04-BP01 Implementar uma política de classificação de dados

Classifique os dados para entender sua importância para os resultados empresariais e selecione o nível de armazenamento eficiente em termos de energia para armazenar os dados.

Antipadrões comuns:

- Você não identifica ativos de dados com características semelhantes (como sensibilidade, importância empresarial ou requisitos regulatórios) que estão sendo processados ou armazenados.

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

Benefícios do estabelecimento desta prática recomendada: a implementação de uma política de classificação de dados permite determinar o nível de armazenamento eficiente em termos de energia para os dados.

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

Orientações para a implementação

A classificação de dados envolve a identificação dos tipos de dados que estão sendo processados e armazenados em um sistema de informação pertencente a uma organização ou operado por ela. Também envolve a decisão em relação à importância dos dados e ao impacto provável do comprometimento dos dados, perda ou uso incorreto.

Implemente a política de classificação de dados trabalhando de forma reversa a partir do uso contextual dos dados e criando um esquema de categorização que leve em conta a importância de determinado conjunto de dados para as operações de uma organização.

Etapas da implementação

- Realize um inventário dos vários tipos de dados que existem para sua workload.
  - Para obter mais detalhes sobre categorias de classificação de dados, consulte o whitepaper Data Classification (Classificação de dados).
- Determine a importância, a confidencialidade, a integridade e a disponibilidade dos dados com base no risco para a organização. Use esses requisitos para agrupar dados em um dos níveis de classificação de dados adotados.
  - Por exemplo, consulte Four simple steps to classify your data and secure your startup (Quatro etapas simples para classificar seus dados e proteger sua startup).
- Audite periodicamente seu ambiente em busca de dados que não estejam etiquetados ou classificados e classifique-os e etiquete-os apropriadamente.
  - Por exemplo, consulte Data Catalog e crawlers no AWS Glue.
- Estabeleça um catálogo de dados que forneça recursos de auditoria e governança.
- Determine e documente procedimentos de manipulação para cada classe de dados.
- Use automação para auditar periodicamente seu ambiente em busca de dados que não estejam etiquetados ou classificados e classifique-os e etiquete-os apropriadamente.

Recursos

Documentos relacionados:

- Leveraging Nuvem AWS to Support Data Classification (Utilizar a Nuvem AWS para oferecer compatibilidade com a classificação de dados)
- Políticas de tag do AWS Organizations

Vídeos relacionados:

- Enabling agility with data governance on AWS (Ativar a agilidade com governança de dados na AWS)

SUS04-BP02 Usar tecnologias compatíveis com seus padrões de acesso e de armazenamento de dados

Use tecnologias de armazenamento mais adequadas à maneira como seus dados são acessados e armazenados a fim de reduzir os recursos provisionados e, ao mesmo tempo, comportar sua workload.

Antipadrões comuns:

- Você pressupõe que todas as workloads têm padrões de acesso e armazenamento de dados semelhantes.
- Você usa apenas um nível de armazenamento, supondo que todas as workloads se encaixem nesse nível.
- Você pressupõe que os padrões de acesso aos dados permanecerão consistentes ao longo do tempo.

Benefícios de estabelecer esta prática recomendada: selecionar e otimizar suas tecnologias de armazenamento com base em padrões de armazenamento e acesso aos dados ajudará a reduzir os recursos de nuvem necessários a fim de atender às suas necessidades empresariais e melhorar a eficiência geral da workload de nuvem.

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

Orientação para implementação

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 eficiência da performance.

- Avalie suas características de dados e padrão de acesso a fim de reunir as principais características de suas necessidades de armazenamento. Principais características a serem consideradas:
  - Tipo de dados: estruturados, semiestruturados e não estruturados
  - Crescimento dos dados: delimitado, não vinculado
  - Durabilidade de dados: persistente, efêmero, transitório
  - Padrões de acesso: leituras ou gravações, frequência, picos ou consistentes
- Migre os dados para a tecnologia de armazenamento apropriada que seja compatível com suas características de dados e padrão de acesso. Veja alguns exemplos de tecnologias de armazenamento da AWS e suas principais características:

Tipo	Tecnologia	Características principais
Armazenamento de objetos	Amazon S3	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 compatibilidade com seu local, necessidades de segurança e padrões de acesso.

Tipo	Tecnologia	Características principais
Armazenamento de arquivamento	Amazon S3 Glacier	Classe de armazenamento do Amazon S3 desenvolvida para arquivamento de dados.
Sistema de arquivos compartilhado	Amazon Elastic File System (Amazon EFS)	Sistema de arquivos montável que pode ser acessado por diversos tipos de soluções de computaçã o. O Amazon EFS aumenta e reduz automaticamente o armazenamento e sua performance é otimizada para oferecer latências baixas de maneira consistente.
Sistema de arquivos compartilhado	Amazon FSx	Baseia-se nas soluções de computação mais recentes da AWS para oferecer compatibi lidade com quatro sistemas de arquivos comumente usados: NetApp ONTAP, OpenZFS, Windows File Server e Lustre. Amazon FSx latência, throughput e IOPS 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.

Tipo	Tecnologia	Características principais
O Armazenamento em bloco	Amazon Elastic Block Store (Amazon EBS)	Serviço de armazenamento de bloco escalável e de alta performance projetado para Amazon Elastic Compute Cloud (Amazon EC2). O Amazon EBS inclui armazenamento com base em SSD para workloads transacionais e de uso intenso de IOPS e armazenamento com base em HDD para workloads de uso intenso de throughput.
Banco de dados relacional	Amazon Aurora, o Amazon RDS, o Amazon Redshift	Projetados para oferecer compatibilidade com transações ACID (atomicidade, consistência, isolamento, durabilidade) e manter a integridade referencial e uma forte consistência de dados. Muitas aplicações tradicionais, sistemas de planejamento de recursos empresariais (ERP), de gerenciamento de relacionamentos com o cliente (CRM) e de comércio eletrônico usam bancos de dados relacionais para armazenar seus dados.

Tipo	Tecnologia	Características principais
Banco de dados de chave-valor	tabelas do Amazon DynamoDB	Otimizados para padrões de acesso comuns, normalmente visando armazenar e recuperar grandes volumes de dados. Aplicações web de alto tráfego, sistemas de comércio eletrônico e aplicações de jogos são os casos de uso habituais para bancos de dados de chave-valor.

- Para sistemas de armazenamento que têm tamanho fixo, como Amazon EBS ou Amazon FSx, monitore o espaço de armazenamento disponível e automatize a alocação de armazenamento ao atingir um limite. Você pode utilizar o Amazon CloudWatch para coletar e analisar diferentes métricas para o Amazon EBS e o Amazon FSx.
- As classes de armazenamento do Amazon S3 podem ser configuradas em nível de objeto e um único bucket pode conter objetos armazenados em todas as classes de armazenamento.
- Você também pode usar políticas de ciclo de vida do Amazon S3 para realizar a transição automática de objetos entre classes de armazenamento ou remover dados sem nenhuma alteração na aplicação. Em geral, você precisa fazer uma compensação entre a eficiência dos recursos, a latência de acesso e a confiabilidade ao considerar esses mecanismos de armazenamento.

Recursos

Documentos relacionados:

- Tipos de volume do Amazon EBS
- Armazenamento de instâncias do Amazon EC2
- Amazon S3 Intelligent-Tiering
- Características de E/S do Amazon EBS
- Uso de classes de armazenamento do Amazon S3
- O que é o Amazon S3 Glacier?

Vídeos relacionados:

- Architectural Patterns for Data Lakes on AWS (Padrões arquitetônicos para data lakes na AWS)
- Deep dive on Amazon EBS (STG303-R1)
- Optimize your storage performance with Amazon S3 (STG343)
- Building modern data architectures on AWS (Criação de arquiteturas de dados na AWS)

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

SUS04-BP03 Usar políticas para gerenciar o ciclo de vida de seus conjuntos de dados

Gerencie o ciclo de vida de todos os seus dados e aplique a exclusão automaticamente para minimizar o armazenamento total necessário para sua workload.

Antipadrões comuns:

- Você exclui dados manualmente.
- Você não exclui nenhum de seus dados de workload.
- Você não move os dados para níveis de armazenamento mais eficientes em termos de energia com base em seus requisitos de retenção e acesso.

Benefícios de estabelecer esta prática recomendada: O uso de políticas de ciclo de vida de dados garante acesso e retenção de dados eficientes em uma workload.

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

Orientações para a implementação

Os conjuntos de dados geralmente têm diferentes requisitos de retenção e acesso durante seu ciclo de vida. Por exemplo, seu aplicativo pode precisar de acesso frequente a alguns conjuntos de

dados por um período limitado. Depois disso, esses conjuntos de dados são acessados com pouca frequência.

Para gerenciar com eficiência seus conjuntos de dados ao longo de seu ciclo de vida, configure políticas de ciclo de vida, que são regras que definem como lidar com conjuntos de dados.

Com as regras de configuração do ciclo de vida, é possível orientar o serviço de armazenamento específico a fazer a transição de um conjunto de dados para níveis de armazenamento mais eficientes em termos de energia, arquivá-lo ou excluí-lo.

Etapas da implementação

- Classifique conjuntos de dados em sua workload.
- Defina procedimentos de manipulação para cada classe de dados.
- Defina políticas automatizadas de ciclo de vida para aplicar regras de ciclo de vida. Aqui estão alguns exemplos de como configurar políticas de ciclo de vida automatizadas para diferentes serviços de armazenamento do AWS:

Storage service	How to set automated lifecycle policies
Amazon S3	Você pode usar o ciclo de vida do Amazon S3 para gerenciar seus objetos durante todo o ciclo de vida. Se seus padrões de acesso forem desconhecidos, mutáveis ou imprevisíveis, você pode usar o Amazon S3 Intelligent-Tiering, que monitora os padrões de acesso e move automaticamente os objetos que não foram acessados para níveis de acesso de custo mais baixo. Você pode aproveitar as métricas da Lente de Armazenamento do Amazon S3 para identificar oportunidades de otimização e lacunas no gerenciamento do ciclo de vida.
Amazon Elastic Block Store	Você pode usar o Amazon Data Lifecycle Manager para automatizar a criação, retenção e exclusão de snapshots do Amazon EBS e AMIs com suporte do Amazon EBS.

Storage service	How to set automated lifecycle policies
Amazon Elastic File System	O gerenciamento do ciclo de vida do Amazon EFS gerencia automaticamente o armazenamento de arquivos para seus sistemas de arquivos.
Amazon Elastic Container Registry	As políticas de ciclo de vida do Amazon ECR automatizam a limpeza de suas imagens de contêiner, expirando imagens com base na idade ou contagem.
AWS Elemental MediaStore	Você pode usar uma política de ciclo de vida do objeto que rege por quanto tempo os objetos devem ser armazenados no contêiner do MediaStore.

- Exclua volumes não utilizados, snapshots e dados que estão fora do período de retenção. Aproveite os recursos do serviço nativo, como o tempo de vida útil do Amazon DynamoDB ou retenção de log do Amazon CloudWatch para exclusão.
- Agregue e compacte dados quando possível com base nas regras do ciclo de vida.

Recursos

Documentos relacionados:

- Otimize suas regras de ciclo de vida do Amazon S3 com a análise de classe de armazenamento do Amazon S3
- Avaliar recursos com o Regras do AWS Config

Vídeos relacionados:

- Simplifique o ciclo de vida de seus dados e otimize os custos de armazenamento com o ciclo de vida do Amazon S3
- Reduza seus custos de armazenamento usando Lente de Armazenamento do Amazon S3

SUS04-BP04 Usar elasticidade e automação para expandir o armazenamento em bloco ou o sistema de arquivos

Use a elasticidade e automação para expandir o armazenamento em bloco ou o sistema de arquivos à medida que os dados aumentarem e minimizar o total de armazenamento provisionado.

Antipadrões comuns:

- Você adquire um grande armazenamento em bloco ou sistema de arquivos para necessidades futuras.
- Você provisiona em excesso as operações de entrada e saída por segundo (IOPS) de seu sistema de arquivos.
- Você não monitora a utilização de seus volumes de dados.

Benefícios do estabelecimento desta prática recomendada: minimizar o provisionamento em excesso do sistema de armazenamento reduz os recursos ociosos e melhora a eficiência geral da workload.

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

Orientações para a implementação

Crie armazenamento em bloco ou sistemas de arquivos com alocação de tamanho, throughput e latência apropriados à workload. Use a elasticidade e automação para expandir o armazenamento em bloco ou o sistema de arquivos à medida que os dados aumentarem sem precisar provisionar em excesso esses serviços de armazenamento.

Etapas da implementação

- Para armazenamento fixo como o Amazon EBS, verifique se está 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.
- 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. A título de exemplo, você pode usar os Volumes Elásticos do Amazon EBS para alterar o tamanho ou o tipo de volume ou ajustar a performance dos volumes do Amazon EBS.
- Escolha a classe de armazenamento, modo de performance e modo de throughput corretos para seu sistema de arquivos para atender à necessidade de seus negócios, sem a ultrapassar.
  - Desempenho do Amazon EFS

  - Performance do volume do Amazon EBS em instâncias Linux
- Defina os níveis pretendidos de utilização para seus volumes de dados e redimensione os volumes fora dos intervalos esperados.
- Dimensione adequadamente 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.
- Analise regularmente volumes elásticos e sistemas de arquivos para encerrar volumes ociosos, reduzir recursos com excesso de provisionamento e se ajustar ao tamanho de dados atual.

Recursos

Documentos relacionados:

- Documentação do Amazon FSx
- O que é o Amazon Elastic File System?

Vídeos relacionados:

- Aprofundamento em Volumes Elásticos do Amazon EBS
- Amazon EBS e estratégias de otimização de snapshots para melhorar a performance e reduzir os custos
- Otimização de custo e performance do Amazon EFS usando práticas recomendadas

SUS04-BP05 Remover dados desnecessários ou redundantes

Remova dados desnecessários ou redundantes para minimizar os recursos de armazenamento necessários para armazenar seus conjuntos de dados.

Antipadrões comuns:

- Você duplica dados que podem ser facilmente obtidos ou recriados.
- Você faz backup de todos os dados sem considerar sua criticidade.
- Você apenas exclui dados irregularmente, em eventos operacionais ou não os exclui.
- Você armazena dados de forma redundante, independentemente da durabilidade do serviço de armazenamento.

- Você ativa o versionamento do Amazon S3 sem qualquer justificativa comercial.

Benefícios do estabelecimento desta prática recomendada: A remoção de dados desnecessários reduz o tamanho de armazenamento necessário para sua workload e o impacto ambiental da workload.

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

Orientações para a implementação

Não armazene dados de que você não precisa. Automatize a exclusão de dados desnecessários. Use tecnologias que eliminem dados duplicados em níveis de arquivo e bloco. Aproveite a replicação de dados nativos e os recursos de redundância dos serviços.

Etapas da implementação

- Avalie se você pode evitar o armazenamento de dados usando conjuntos de dados disponíveis publicamente no AWS Data Exchange e Dados abertos no AWS.
- Use mecanismos que possam duplicar dados no nível de bloco e objeto. Aqui estão alguns exemplos de como desduplicar dados no AWS:

Storage service	Deduplication mechanism
Amazon S3	Use AWS Lake Formation FindMatches para localizar registros correspondentes em um conjunto de dados (incluindo aqueles sem identificadores) usando o novo FindMatches ML Transform.
Amazon FSx	Habilite a desduplicação de dados no Amazon FSx para Windows.
Snapshots do Amazon Elastic Block Store	Os snapshots são backups incrementais, o que significa que apenas os blocos no dispositivo que foram alterados após o snapshot mais recente são salvos.

- Analise o acesso aos dados para identificar dados desnecessários. Automatize as políticas de ciclo de vida. Aproveite os recursos do serviço nativo, como o tempo de vida útil do Amazon

DynamoDB, ciclo de vida do Amazon S3 ou retenção de log do Amazon CloudWatch para exclusão.

- Use os recursos de virtualização de dados no AWS para manter os dados em sua origem e evitar a duplicação de dados.
  - Virtualização de dados nativos da nuvem no AWS
  - Laboratório: Otimizar padrão de dados usando o compartilhamento de dados do Amazon Redshift
- Use a tecnologia de backup que pode fazer backups incrementais.
- Aproveite a durabilidade do Amazon S3 e a replicação do Amazon EBS para atender às suas metas de durabilidade em vez de tecnologias autogerenciadas (como uma Redundant Array of Independent Disks [RAID – Matriz redundante de discos independentes]).
- 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:

- 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

Vídeos relacionados:

- Correspondência difusa e desduplicação de dados com ML Transforms para o AWS Lake Formation

Exemplos relacionados:

- Como analiso meus logs de acesso ao servidor do Amazon S3 usando o Amazon Athena?

SUS04-BP06 Usar sistemas de arquivos compartilhados ou armazenamento para acessar dados comuns

Adote armazenamento ou sistemas de arquivos compartilhados para evitar a duplicação de dados e viabilize fuma infraestrutura mais eficiente para sua workload.

Antipadrões comuns:

- Você provisiona armazenamento para cada cliente específico.
- Você não desanexa o volume de dados dos clientes inativos.
- Você não fornece acesso a armazenamento em plataformas e sistemas.

Benefícios do estabelecimento desta prática recomendada: o uso de armazenamento ou sistemas de arquivos permite que os dados sejam compartilhados com um ou mais consumidores sem precisar copiá-los. Isso ajuda a reduzir os recursos de armazenamento necessários à workload.

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

Orientações para a implementação

Se você tiver vários usuários ou aplicações que acessam os mesmos conjuntos de dados, o uso da tecnologia de armazenamento compartilhado é essencial para viabilizar uma infraestrutura eficiente para sua workload. A tecnologia de armazenamento compartilhado oferece um local central para armazenar e gerenciar conjuntos de dados e evitar a duplicação de dados. Ela também impõe a consistência dos dados em sistemas diferentes. Além disso, a tecnologia de armazenamento compartilhado permite o uso mais eficiente da potência computacional, visto que vários recursos podem acessar e processar os dados ao mesmo tempo em paralelo.

Busque dados dos serviços de armazenamento compartilhado somente quando necessário e desanexe os volumes não usados para liberar recursos.

Etapas da implementação

- Migre dados para o armazenamento compartilhado quando eles tiverem vários consumidores. Veja aqui alguns exemplos de tecnologia de armazenamento compartilhado na AWS:

Storage option	When to use
Amazon EBS Multi-Attach	O Amazon EBS Multi-Attach permite que você anexe um único volume de SSD de IOPS provisionados (io1 ou io2) para várias instâncias que estão na mesma zona de disponibilidade.
Amazon EFS	Consulte Quando escolher o Amazon EFS.
Amazon FSx	Consulte Como escolher um Amazon FSx File System).
Amazon S3	As aplicações que não exigem uma estrutura de sistema de arquivos e são projetadas para funcionar com armazenamento de objetos podem usar o Amazon S3 como solução de armazenamento de objetos altamente escalável, durável e de baixo custo.

- Copie os dados para ou busque os dados de sistemas de arquivos compartilhados somente quando necessário. A título de exemplo, você pode usar um sistema de arquivos do Amazon FSx for Lustre respaldado pelo Amazon S3 e carregar somente o subconjunto de dados necessário para processar os trabalhos para o Amazon FSx.
- Exclua dados conforme apropriado para os seus padrões de uso, tal como descrito em SUS04-BP03 Usar políticas para gerenciar o ciclo de vida de seus conjuntos de dados.
- Desvincule volumes de clientes que não estão utilizando-os ativamente.

Recursos

Documentos relacionados:

- Vincular seu sistema de arquivos a um bucket do Amazon S3
- Como usar o Amazon EFS para AWS Lambda em suas aplicações sem servidor
- Amazon EFS Intelligent-Tiering Intelligent-Tiering otimiza os custos para workloads com padrões de acesso variáveis

- Como usar o Amazon FSx com seu repositório de dados on-premises

Vídeos relacionados:

- Otimização de custos de armazenamento com o Amazon EFS

SUS04-BP07 Minimizar a movimentação de dados entre redes

Use o armazenamento de objetos ou os sistemas de arquivos compartilhados para acessar dados comuns e minimizar os recursos totais de rede exigidos para comportar a movimentação de dados da workload.

Antipadrões comuns:

- Você armazena todos os dados na mesma Região da AWS independentemente de onde os usuários dos dados estão.
- Você não otimiza o tamanho e o formato dos dados antes de movimentá-los na rede.

Benefícios de estabelecer esta prática recomendada: otimizar a movimentação de dados na rede reduz os recursos totais de rede necessários à workload e reduz o respectivo impacto ambiental.

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

Orientação para implementação

A movimentação de dados em sua organização exige recursos de computação, rede e armazenamento. Use técnicas para minimizar a movimentação de dados e melhorar a eficiência geral da workload.

Etapas da implementação

- Considere a proximidade dos dados ou dos usuários como um fator de decisão ao selecionar uma região para sua workload.
- Particione serviços consumidos regionalmente para que os dados específicos da região sejam armazenados na região em que eles são consumidos.
- Use formatos de arquivo eficientes (como Parquet ou ORC) e compacte os dados antes movimentá-los na rede.

- Não movimente dados não usados. Alguns exemplos que podem ajudar você a evitar a movimentação de dados não utilizados:
  - Reduza as respostas de API apenas aos dados relevantes.
  - Agregue os dados onde não houver necessidade de informações detalhadas (em nível de registro).
  - Consulte Laboratório do Well-Architected: Otimizar padrão de dados usando o compartilhamento de dados do Amazon Redshift..
  - Considere o compartilhamento de dados entre contas no AWS Lake Formation.
- Use serviços que podem ajudar você a executar código mais perto dos usuários da workload.

Service	Quando usar
o Lambda@Edge	Use para operações com uso intenso de computação que são executadas quando objetos não estão no cache.
Funções do CloudFront	Use para casos de uso simples como solicitações HTTP(s)/manipulações de resposta que podem ser iniciadas por funções de curta duração.
AWS IoT Greengrass	Execute computação local, mensagens e armazenamento de dados em cache para dispositivos conectados.

Recursos

Documentos relacionados:

- 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

Vídeos relacionados:

- Demystifying data transfer on AWS (Desmistificação da transferência de dados na AWS)

Exemplos relacionados:

- Arquitetura para a sustentabilidade: reduza a movimentação de dados entre redes

SUS04-BP08 Fazer backup de dados somente quando for difícil recriar

Evite fazer backup de dados que não têm valor empresarial para minimizar os requisitos de recurso de armazenamento da workload.

Antipadrões comuns:

- Você não tem uma estratégia de backup para seus dados.
- Você faz backup de dados que podem ser recriados com facilidade.

Benefícios do estabelecimento desta prática recomendada: evitar backups de dados não essenciais reduz os recursos de armazenamento necessários à workload e diminui o respectivo impacto ambiental.

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

Orientações para a implementação

Evitar o backup de dados desnecessários pode ajudar a reduzir os custos e os recursos de armazenamento usados pela workload. 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.

Etapas da implementação

- Implemente uma política de classificação de dados como descrito em SUS04-BP01 Implementar uma política de classificação de dados.

- Use a criticidade da estratégia de classificação de dados e backup de design com base no objetivo de tempo de recuperação (RTO) e no objetivo de ponto de recuperação (RPO). Evite fazer backup de dados não essenciais.
  - Exclua os dados que podem ser recriados com facilidade.
  - 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).
- Use uma solução automatizada ou um serviço gerenciado para fazer backup de dados essenciais aos negócios.
  - O AWS Backup é um serviço totalmente gerenciado que facilita a centralização e a automatização da proteção de dados nos serviços da AWS, na nuvem e no ambiente on-premises. Para obter orientações práticas sobre como criar backups automatizados usando o AWS Backup, consulte Laboratórios do Well-Architected: Teste de backup e restauração de dados.
  - Automatize backups e otimize os custos de backup do Amazon EFS usando o AWS Backup.

Recursos

Práticas recomendadas relacionadas:

- REL09-BP01 Identificar e fazer backup de todos os dados que precisam de backup ou reproduzir os dados das fontes
- REL09-BP03 Realizar o backup de dados automaticamente
- REL13-BP02 Usar estratégias de recuperação definidas para atender aos objetivos de recuperação

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
- Parceiro do APN: parceiros que podem ajudar com o backup
- AWS Marketplace: produtos que podem ser usados para backup
- Como fazer backup do Amazon EFS

- Como fazer backup do Amazon FSx para Windows File Server
- Backup e restauração do Amazon ElastiCache for Redis

Vídeos relacionados:

- AWS re:Invent 2021: Backup, recuperação de desastres e proteção contra ransomware com a AWS
- AWS Backup Demonstração do AWS Backup: Backup entre contas e entre regiões
- 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
- Laboratório do Well-Architected: Backup e restauração com failback para workload do Analytics
- Laboratório do Well-Architected: Recuperação de desastres: backup e restauração

Hardware e serviços

Pergunta

- SUS 5 Como selecionar e usar hardware e serviços em nuvem na arquitetura para apoiar os objetivos de sustentabilidade?

SUS 5 Como selecionar e usar hardware e serviços em nuvem na arquitetura para apoiar os objetivos 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 e os serviços mais eficientes para sua workload específica.

Práticas recomendadas

- SUS05-BP01 Usar a quantidade mínima de hardware para atender às suas necessidades
- SUS05-BP02 Usar tipos de instância com o mínimo de impacto
- SUS05-BP03 Usar serviços gerenciados

- SUS05-BP04 Otimizar o uso de aceleradores de computação baseados em hardware

SUS05-BP01 Usar a quantidade mínima de hardware para atender às suas necessidades

Use a quantidade mínima de hardware para sua workload para atender com eficiência às suas necessidades de negócios.

Antipadrões comuns:

- Você não monitora a utilização de recursos.
- Você tem recursos com baixo nível de utilização em sua arquitetura.
- Você não analisa a utilização de hardware estático para determinar se é necessário redimensioná-lo.
- Você não define metas de utilização de hardware para sua estrutura de computação com base nos KPIs de negócios.

Benefícios do estabelecimento desta prática recomendada: dimensionar corretamente os recursos de nuvem ajuda a reduzir o impacto ambiental da workload, a economizar e a manter os parâmetros de performance.

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

Orientações para a implementação

Selecione do modo mais eficiente o número total de hardware necessário à workload para melhorar a eficiência geral. A Nuvem AWS oferece flexibilidade para expandir ou reduzir dinamicamente a quantidade de recursos por meio de diversos mecanismos, como o AWS Auto Scaling, para atender a mudanças na demanda. Ela também fornece APIs e SDKs que permitem que os recursos sejam modificados com o mínimo de esforço. Use esses recursos para fazer alterações frequentes nas implementações da workload. Além disso, use as orientações sobre dimensionamento correto das ferramentas da AWS para operar com eficiência o recursos de nuvem e atender às suas necessidades empresariais.

Etapas da implementação

- Escolha os tipos de instância mais adequados às suas necessidades.
  - Como faço para escolher o tipo de instância do Amazon EC2 apropriado para minha workload?

  - Seleção de tipo de instância baseada em atributos para frota do Amazon EC2.
  - Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos.
- Escale usando pequenos incrementos para workloads variáveis.
- Use várias opções de compra de computação para equilibrar flexibilidade, escalabilidade e economia no uso de instâncias.
  - As instâncias sob demanda são mais adequadas para workloads novas, com estado e com picos que não podem ter flexibilidade com relação ao tipo de instância, local e tempo.
  - As instâncias spot são excelentes para complementar as outras opções para aplicações tolerantes a falhas e flexíveis.
  - Utilize Savings Plans para computação para workloads de estado estável que permitem flexibilidade se suas necessidades (como AZ, região, famílias de instâncias ou tipos de instância) mudarem.
- Use uma variedade de instâncias e zonas de disponibilidade para maximizar a disponibilidade da aplicação e aproveitar o excesso de capacidade quando possível.
- Use as recomendações de dimensionamento correto das ferramentas da AWS para fazer ajustes na workload.
  - AWS Compute Optimizer
  - AWS Trusted Advisor
- Negocie Acordos de Serviço (SLAs) que permitam uma redução temporária na capacidade enquanto a automação implanta recursos de substituição.

Recursos

Documentos relacionados:

- Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte I: Computação
- Seleção de tipo de instância baseada em atributo do Auto Scaling para Amazon EC2 Fleet
- Documentação do AWS Compute Optimizer
- Otimização do Lambda: otimização da performance
- Documentação do Auto Scaling

Vídeos relacionados:

- Criar um ambiente de computação eficiente em termos de custo, energia e recursos

Exemplos relacionados:

- Laboratórios do Well-Architected: Dimensionamento correto com o AWS Compute Optimizer e utilização de memória habilitada (Nível 200)

SUS05-BP02 Usar tipos de instância com o mínimo de impacto

Monitore continuamente e use novos tipos de instância para aproveitar as melhorias de eficiência de energia.

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: Médio

Orientação para implementação

Usar instâncias eficientes na workload de nuvem é essencial para reduzir o uso de recursos e os custos. 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 e inferência de machine learning e transcodificação de vídeo.

Etapas da 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.
    - AWS Graviton Fast Start
    - Considerações ao migrar workloads para instâncias do Amazon Elastic Compute Cloud baseadas no AWS Graviton
    - AWS Graviton2 para ISVs
  - 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, utilize hardware específico para sua workload, como AWS Trainium, AWS Inferentiae aos Amazon EC2 DL1. Instâncias do AWS Inferentia, como instâncias Inf2, oferecem performance até 50% melhor por watt em relação a instâncias comparáveis do Amazon EC2.
  - Use o Amazon SageMaker Inference Recommender para dimensionar endpoints de inferência de ML corretamente.
  - Para workloads com picos (workloads com requisitos irregulares 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.
        - Laboratório do Well-Architected: Recomendações de dimensionamento correto
        - Laboratório do Well-Architected: Dimensionamento correto com o Compute Optimizer
        - Laboratório do Well-Architected: Otimizar padrões de hardware e observar KPIs de sustentabilidade

Recursos

Documentos relacionados:

- Otimizar a sua infraestrutura da AWS para sustentabilidade, Parte I: Computação
- AWS Graviton
- Amazon EC2 DL1
- Frotas de reserva de capacidade do Amazon EC2
- Frota spot do Amazon EC2
- Funções: configuração de função do Lambda
- Seleção de tipo de instância baseada em atributos para frota do Amazon EC2
- Criar aplicações sustentáveis, eficientes e com custo otimizado na AWS
- Como o Painel de Sustentabilidade da Contino ajuda os clientes a otimizar sua pegada de carbono

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)

- Criar um ambiente de computação eficiente em termos de custo, energia e recursos

Exemplos relacionados:

- Solução: orientações sobre como otimizar workloads de aprendizado profundo para sustentabilidade na AWS
- Laboratório do Well-Architected: Recomendações de dimensionamento correto
- Laboratório do Well-Architected: Dimensionamento correto com o Compute Optimizer
- Laboratório do Well-Architected: Otimizar padrões de hardware e observar KPIs de sustentabilidade
- Laboratório do Well-Architected: Migração de serviços para o Graviton

SUS05-BP03 Usar serviços gerenciados

Use serviços gerenciados para operar com maior eficiência na nuvem.

Antipadrões comuns:

- Você usa instâncias do Amazon EC2 com baixa utilização para executar suas aplicações.
- Sua equipe interna gerencia apenas a workload e não tem tempo para se concentrar em inovação ou simplificações.
- Você implanta e mantém tecnologias para tarefas que podem ser executadas com maior eficiência em serviços gerenciados.

Benefícios do estabelecimento desta prática recomendada:

- Com o uso de serviços gerenciados, a responsabilidade é transferida para a AWS, que tem insights referentes a milhões de clientes que podem ajudar a promover inovações inéditas e melhorar a eficiência.
- O serviço gerenciado distribui o impacto ambiental do serviço entre vários usuários por causa do ambiente de gerenciamento de vários locatários.

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

Orientações para a implementação

Como os serviços gerenciados, a responsabilidade por manter a alta utilização e otimizar a sustentabilidade do hardware implantado é transferida para a AWS. Os serviços gerenciados também eliminam as despesas operacionais e administrativas da manutenção de um serviço, o que permite que sua equipe tenha mais tempo para se concentrar na inovação.

Avalie sua workload para identificar componentes que podem ser substituídos por serviços gerenciados da AWS. Por exemplo, o Amazon RDS, o Amazon Redshift e o Amazon ElastiCache fornecem um serviço gerenciado de banco de dados. O Amazon Athena, o Amazon EMR e o Amazon OpenSearch Service fornecem um serviço gerenciado de análise.

Etapas da implementação

1. Faça um inventário de serviços e componentes para sua workload.
2. Avalie e identifique componentes que podem ser substituídos por serviços gerenciados. Veja aqui alguns exemplos de quando você pode pensar em usar um serviço gerenciado:

Task	What to use on AWS
Hospedagem de banco de dados	Use instâncias do Amazon Relational Database Service (Amazon RDS) gerenciadas, em vez de manter suas próprias instâncias do Amazon RDS no Amazon Elastic Compute Cloud (Amazon EC2).
Hospedagem de uma workload conteinerizada	Use o AWS Fargate, em vez de implementar sua própria infraestrutura de contêiner.
Hospedagem de aplicações Web	Use o AWS Amplify Hosting como CI/CD totalmente gerenciadas e serviço de hospedagem para sites estáticos e aplicações Web renderizadas do lado do servidor.

3. Identifique dependências e crie um plano de migração. Atualize runbooks e manuais and playbooks de forma apropriada.
   - O AWS Application Discovery Service coleta e apresenta automaticamente informações detalhadas sobre dependências e utilização de aplicações para ajudar você a tomar decisões bem fundamentadas ao planejar a migração.

4. Teste o serviço antes de migrar para o serviço gerenciado.
5. Use o plano de migração para substituir serviços auto-hospedados por serviço gerenciado.
6. Monitore continuamente o serviço após a conclusão da migração para fazer ajustes conforme necessário e otimizar o serviço.

Recursos

Documentos relacionados:

- Produtos da Nuvem AWS
- Calculadora de custo total de propriedade (TCO) da AWS
- Amazon DocumentDB
- Amazon Elastic Kubernetes Service (EKS)
- Amazon Managed Streaming for Apache Kafka (Amazon MSK)

Vídeos relacionados:

- Operações em nuvem em escala com o AWS Managed Services

SUS05-BP04 Otimizar o uso de aceleradores de computação baseados em hardware

Otimize o uso de instâncias com computação acelerada para reduzir as demandas de infraestrutura física de sua workload.

Antipadrões comuns:

- Você não está monitorando o uso da GPU.
- Você está usando uma instância de finalidade geral para workload, enquanto uma instância criada especificamente pode oferecer maior desempenho, menor custo e melhor desempenho por watt.
- Você está usando aceleradores de computação baseados em hardware para tarefas em que são mais eficientes usando alternativas baseadas em CPU.

Benefícios de estabelecer esta prática recomendada: Ao otimizar o uso de aceleradores baseados em hardware, você pode reduzir as demandas de infraestrutura física de sua workload.

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

Orientação para implementação

Se você precisar de alta capacidade de processamento, poderá se beneficiar do uso de instâncias com computação acelerada, que fornecem acesso a aceleradores de computação baseados em hardware, como unidades de processamento gráfico (GPUs) e matrizes de portas programáveis em campo (FPGAs). Esses aceleradores de hardware executam certas funções, como processamento gráfico ou correspondência de padrões de dados, com mais eficiência do que alternativas baseadas em CPU. Muitas workloads aceleradas, como renderização, transcodificação e machine learning, são altamente variáveis em termos de uso de recursos. Execute este hardware somente pelo tempo necessário e desative-as com automação quando não precisar mais delas para reduzir o consumo de recursos.

Etapas da implementação

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

Recursos

Documentos relacionados:

- Computação acelerada

- Vamos, arquiteto! Arquitetura com chips e aceleradores personalizados
- Como faço para escolher o tipo de instância do Amazon EC2 apropriado para minha workload?
- Instâncias VT1 do Amazon EC2
- Escolha o melhor acelerador de IA e compilação de modelo para inferência de visão computacional com o Amazon SageMaker

Vídeos relacionados:

- How to select Amazon EC2 GPU instances for deep learning (Como selecionar instâncias de GPU do Amazon EC2 para aprendizado profundo)
- Deploying Cost-Effective Deep Learning Inference (Implantação de inferência de aprendizado profundo econômico)

Processo e cultura

Pergunta

- SUS 6 Como os processos organizacionais apoiam as metas de sustentabilidade?

SUS 6 Como os processos organizacionais apoiam as 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
- SUS06-BP02 Manter a workload atualizada
- SUS06-BP03 Aumentar a utilização de ambientes de desenvolvimento
- SUS06-BP04 Usar parques de dispositivos gerenciados para testes

SUS06-BP01 Adotar métodos que podem apresentar melhorias na sustentabilidade rapidamente

Adote métodos e processos para validar possíveis aprimoramentos, minimizar o custo dos testes e fornecer pequenas melhorias.

Antipadrões comuns:

- A avaliação da sustentabilidade de sua aplicação é uma tarefa que é feita apenas uma vez no início de um projeto.
- Como o processo de lançamento para introduzir pequenas alterações em prol da eficiência dos recursos é muito trabalhoso, sua workload ficou ultrapassada.
- Você não tem mecanismos para melhorar a sustentabilidade de sua workload.

Benefícios do estabelecimento desta prática recomendada: por meio da criação de um processo para introduzir e monitorar melhorias de sustentabilidade, você conseguirá adotar novos recursos e capacidades, eliminar problemas e melhorar a eficiência da workload continuamente.

Nível de risco exposto se esta prática recomendada não é estabelecida: médio

Orientações para a implementação

Teste e valide as possíveis melhorias de sustentabilidade 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 oferecer pequenas melhorias.

Etapas da implementação

- Adicione requisitos de melhoria da sustentabilidade às suas pendência de desenvolvimento.
- Use um processo de melhoria iterativo para identificar, avaliar, priorizar, testar e implantar essas melhorias.
- Melhore e otimize continuamente seus processos de desenvolvimento. A título de exemplo, automatize seu processo de entrega de software usando pipelines de integração contínua e entrega contínua a fim de testar e implantar possíveis melhorias para reduzir o nível de esforço e limitar os erros provocados por processos manuais.
- Desenvolva e teste possíveis melhorias usando os componentes representativos mínimos viáveis para reduzir o custo dos testes.
- Avalie continuamente o impacto das melhorias e faça ajustes conforme necessário.

Recursos

Documentos relacionados:

- A AWS viabiliza soluções de sustentabilidade

- Práticas de desenvolvimento ágil escaláveis com base no AWS CodeCommit

Vídeos relacionados:

- Entregar arquiteturas sustentáveis e de alta performance

Exemplos relacionados:

- Laboratório do Well-Architected: Transformação de relatório de custo e uso em relatório de eficiência

SUS06-BP02 Manter a workload atualizada

Mantenha sua workload atualizada para adotar recursos eficientes, eliminar problemas e melhorar a eficiência geral da workload.

Antipadrões comuns:

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

Benefícios do estabelecimento desta prática recomendada: ao estabelecer um processo para manter a workload atualizada, você pode adotar novos recursos e capacidades, resolver problemas e aumentar a eficiência da workload.

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

Orientações para a implementação

Sistemas operacionais, tempos de execução, middleware, 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 maior precisão, pois os fornecedores oferecem recursos para atender às suas próprias metas de sustentabilidade. Adote um ritmo regular para manter a workload atualizada com os recursos e versões mais recentes.

Etapas da 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 AWS Systems Manager Inventory para coletar metadados de sistema operacional (SO), aplicação e instância das instâncias do Amazon EC2 e entender rapidamente quais instâncias executam o software e as configurações exigidas pela política de software e quais instâncias precisam ser atualizadas.
- Entenda como atualizar os componentes de sua workload.

Workload component	How to update
Imagens de máquina	Use o EC2 Image Builder para gerenciar atualizações de imagens de máquina da Amazon (AMIs) para imagens de servidor Linux ou Windows.
Imagens de contêiner	Use o Amazon Elastic Container Registry (Amazon ECR) com seu pipeline para gerenciar Amazon Elastic Container Service (Amazon ECS) imagens.
AWS Lambda	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.
  - Você pode usar CI/CD para atualizar automaticamente AMIs, imagens de contêiner e outros artefatos relacionados à sua aplicação de nuvem.

- Você pode usar ferramentas como o Gerenciador de Patches do AWS Systems Manager para automatizar o processo de atualizações de sistema e programar a atividade usando as Janelas de Manutenção do AWS Systems Manager.

Recursos

Documentos relacionados:

- Central de arquitetura da AWS
- Quais as novidades da AWS
- Ferramentas do desenvolvedor na AWS

Exemplos relacionados:

- Laboratórios do Well-Architected: Gerenciamento de inventário e patches
- Laboratório: AWS Systems Manager

SUS06-BP03 Aumentar a utilização de ambientes de desenvolvimento

Aumente a utilização dos recursos para desenvolver, testar e compilar suas workloads.

Antipadrões comuns:

- Você provisiona ou encerra manualmente seus ambientes de compilação.
- Você mantém seus ambientes de compilação em execução independentemente de atividades de teste, compilação ou lançamento (por exemplo, execução de um ambiente fora do horário de expediente dos membros de sua equipe de desenvolvimento).
- Você provisiona recursos em excesso para seus ambientes de compilação.

Benefícios do estabelecimento desta prática recomendada: ao aumentar a utilização dos ambientes de compilação, você pode melhorar a eficiência geral de sua workload de nuvem e, ao mesmo tempo, alocar recursos aos compiladores para que eles desenvolvam, testem e compilem com eficiência.

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

Orientações para a implementação

Use a automação e a infraestrutura como código para ativar ambientes de compilação 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 configuração dos ambientes de teste deve ser bem semelhante à do ambiente de produção. Entretanto, procure oportunidades para usar tipos de instância com capacidade de expansão, instâncias spot do Amazon EC2, serviços de banco de dados com escalabilidade automática, contêineres e tecnologias sem servidor para alinhar a capacidade de desenvolvimento e teste com o uso. Limite o volume de dados apenas para atender os requisitos de teste. Se usar dados de produção no teste, explore possibilidades para compartilhar os dados da produção em vez de movimentá-los.

Etapas da implementação

- Use a infraestrutura como código para provisionar os ambientes de compilação.
- Use a automação para gerenciar o ciclo de vida de seus ambientes de desenvolvimento e teste e maximizar a eficiência dos recursos de compilação.
- Use estratégias para maximizar a utilização de seus ambientes de desenvolvimento e teste.
  - Use ambientes representativos mínimos viáveis para desenvolver e testar possíveis melhorias.
  - Utilize tecnologias sem servidor, se possível.
  - Use instâncias sob demanda para complementar os dispositivos de desenvolvedor.
  - 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.
  - Escale automaticamente seus recursos de compilação de acordo com seus trabalhos de compilação.

Recursos

Documentos relacionados:

- Gerenciador de sessões do AWS Systems Manager
- Instâncias de performance expansível do Amazon EC2
- O que é o AWS CloudFormation?

- O que é o AWS CodeBuild?
- Programador de Instâncias da AWS

Vídeos relacionados:

- Práticas recomendadas de integração contínua

SUS06-BP04 Usar parques de dispositivos gerenciados para testes

Use parques de dispositivos gerenciados para testar com eficiência um novo recurso em um conjunto representativo de hardware.

Antipadrões comuns:

- Você testa e implanta manualmente sua aplicação em dispositivos físicos individuais.
- Você não usa o serviço de testes de aplicação para testar e interagir com suas aplicações (por exemplo, Android, iOS e aplicações Web) em dispositivos físicos reais.

Benefícios do estabelecimento desta prática recomendada: usar parques de dispositivos gerenciados para testar aplicações habilitadas para a nuvem oferece inúmeros benefícios:

- Eles contam com recursos mais eficientes para testar a aplicação em uma ampla variedade de dispositivos.
- Eles eliminam a necessidade de infraestrutura interna para testes.
- Eles oferecem diversos tipos de dispositivo, incluindo hardware mais antigo e menos conhecido, eliminando a necessidade de atualizações de dispositivo desnecessárias.

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

Orientações para a implementação

Usar parques de dispositivos gerenciados pode ajudar a otimizar o processo de testes de novos recursos em um conjunto representativo de hardware. Os parques de dispositivos gerenciados oferecem diversos tipos de dispositivo, incluindo hardware mais antigo e menos conhecido, e evita o impacto sobre a sustentabilidade por parte do cliente devido a atualizações desnecessárias de dispositivo.

Etapas da implementação

- Defina seus requisitos e plano de testes (como tipo de teste, sistemas operacionais e programação dos testes).
  - Você pode usar o Amazon CloudWatch RUM para coletar e analisar dados do lado do cliente e moldar seu plano de testes.
- Selecione o parque de dispositivos gerenciados capaz de atender aos seus requisitos de teste. Por exemplo, você pode usar o AWS Device Farm para testar e conhecer o impacto de suas alterações sobre um conjunto representativo de hardware.
- Use a integração contínua/implantação contínua (CI/CD) para programar e executar seus testes.
  - Integração do AWS Device Farm com seu pipeline de CI/CD para executar testes Selenium entre navegadores
  - Compilação e teste de aplicativos iOS e iPadOS com DevOps e serviços móveis da AWS
- Avalie continuamente os resultados dos testes e faça as melhorias necessárias.

Recursos

Documentos relacionados:

- Lista de dispositivos do AWS Device Farm
- Visualização do painel do CloudWatch RUM

Exemplos relacionados:

- Aplicativo de exemplo do AWS Device Farm para Android
- Aplicativo de exemplo do AWS Device Farm para iOS
- estes Web Appium para AWS Device Farm

Vídeos relacionados:

- Otimização de aplicações por meio de insights sobre o usuário final com o Amazon CloudWatch RUM

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.