AWS Fundamentos de várias regiões

AWS Orientação prescritiva

AWS Orientação prescritiva: AWS Fundamentos de várias regiões

Table of Contents

AWS Fundamentos de várias regiões

John Formento, Amazon Web Services ()AWS

Setembro de 2025 (histórico do documento)

Este guia avançado de 300 níveis é destinado a arquitetos de nuvem e líderes seniores que criam cargas de trabalho AWS e estão interessados em usar uma arquitetura multirregional para melhorar a resiliência de suas cargas de trabalho. Este guia pressupõe um conhecimento básico de AWS infraestrutura e serviços. Ele descreve casos de uso comuns de várias regiões, compartilha conceitos e implicações fundamentais de várias regiões sobre design, desenvolvimento e implantação e fornece orientação prescritiva para ajudá-lo a determinar melhor se uma arquitetura multirregional é adequada para suas cargas de trabalho.

Neste guia:

Você é Well-Architected?

O AWS Well-Architected Framework ajuda você a entender os prós e os contras das decisões que você toma ao criar sistemas na nuvem. Os seis pilares da Estrutura fornecem as melhores práticas arquitetônicas para projetar e operar sistemas confiáveis, seguros, eficientes, econômicos e sustentáveis. Você pode usar o AWS Well-Architected Tool, que está disponível gratuitamente no Console de gerenciamento da AWS, para analisar suas cargas de trabalho em relação a essas melhores práticas, respondendo a um conjunto de perguntas para cada pilar.

Para obter orientação especializada adicional e melhores práticas para sua arquitetura de nuvem, incluindo implantações de arquitetura de referência, diagramas e guias técnicos, consulte o AWS Architecture Center.

Introdução

Cada uma Região da AWSconsiste em várias zonas de disponibilidade independentes e fisicamente separadas dentro de uma área geográfica. A separação lógica estrita entre os serviços de software em cada região é mantida. Esse design proposital garante que uma falha de infraestrutura ou serviço em uma região não resulte em uma falha correlacionada em outra região.

A maioria dos AWS usuários pode atingir seus objetivos de resiliência para uma carga de trabalho em uma única região usando várias zonas de disponibilidade ou regionais. Serviços da AWS No entanto, um subconjunto de usuários busca arquiteturas multirregionais por três motivos:

- Eles têm requisitos de alta disponibilidade e continuidade de operações para suas cargas de trabalho de nível mais alto e desejam estabelecer um tempo de recuperação limitado contra deficiências que afetam os recursos em uma única região.
- Eles precisam atender aos requisitos de soberania de dados (como a adesão às leis, regulamentações e conformidade locais) que exigem que as cargas de trabalho operem dentro de uma determinada jurisdição.
- Eles precisam melhorar o desempenho e a experiência do cliente para a carga de trabalho executando as cargas de trabalho em locais mais próximos de seus usuários finais.

Este guia se concentra nos requisitos de alta disponibilidade e continuidade das operações e ajuda você a abordar as considerações sobre a adoção de uma arquitetura multirregional para uma carga de trabalho. Ele descreve conceitos fundamentais que se aplicam ao design, desenvolvimento e implantação de uma carga de trabalho multirregional e fornece uma estrutura prescritiva para ajudá-lo a determinar se uma arquitetura multirregional é a escolha certa para uma carga de trabalho específica. Você precisa garantir que uma arquitetura multirregional seja a escolha certa para sua carga de trabalho, pois essas arquiteturas são desafiadoras e, se a arquitetura multirregional não for criada corretamente, é possível que a disponibilidade geral da carga de trabalho diminua.

Engenharia e operação para resiliência em uma única região

Antes de mergulhar nos conceitos de várias regiões, comece confirmando que sua carga de trabalho já é a mais resiliente possível em uma única região. Para conseguir isso, avalie sua carga de trabalho em relação ao pilar de confiabilidade e excelência operacional do AWS Well-Architected Framework e faça as alterações necessárias com base em compensações e avaliação de riscos. Os seguintes conceitos são abordados no AWS Well-Architected Framework:

- Segmentação da carga de trabalho com base nos limites do domínio
- Contratos de serviço bem definidos
- Gerenciamento de dependências e acoplamento
- Lidando com falhas, novas tentativas e estratégias de recuo
- Operações idempotentes e transações com estado versus transações sem estado
- Prontidão operacional e gerenciamento de mudanças
- Entendendo a integridade da carga de trabalho
- Respondendo a eventos

Para aprofundar a resiliência em uma única região, revise e aplique os conceitos discutidos no paper Advanced Multi-AZ Resilience Patterns: Detecting and Mitigating Gray Failures. Este paper fornece as melhores práticas para o uso de réplicas em cada zona de disponibilidade para conter falhas e expande os conceitos Multi-AZ que são introduzidos no AWS Well Architected Framework. Embora uma arquitetura multirregional possa mitigar os modos de falha vinculados às zonas de disponibilidade, há vantagens e desvantagens decorrentes de uma abordagem multirregional que você deve considerar. É por isso que recomendamos que você comece com uma abordagem Multi-AZ e, em seguida, avalie uma carga de trabalho específica em relação aos fundamentos de arquiteturas multirregionais para determinar se uma abordagem multirregional pode aumentar a resiliência da carga de trabalho.

Princípio fundamental da multirregião 1: Compreender os requisitos

Conforme mencionado anteriormente, a alta disponibilidade e a continuidade das operações são motivos comuns para buscar arquiteturas multirregionais. As métricas de disponibilidade medem a porcentagem de tempo em que uma carga de trabalho está disponível para uso em um período definido, enquanto as métricas de continuidade das operações medem o tempo de recuperação de eventos de grande escala e, normalmente, de maior duração.

Medir a disponibilidade é um processo quase contínuo. As medições específicas podem variar, mas normalmente se aglutinam em torno de uma métrica de disponibilidade alvo, geralmente chamada de nove (como disponibilidade de 99,99%). Com metas de disponibilidade, um tamanho único não serve para todos. Você deve estabelecer metas de disponibilidade no nível da carga de trabalho e separar os componentes não críticos dos componentes críticos, em vez de aplicar uma única meta em todas as cargas de trabalho.

Para a continuidade das operações, as seguintes point-in-time medidas são normalmente usadas:

- Objetivo de tempo de recuperação (RTO) — RTO é o atraso máximo aceitável entre a interrupção do serviço e a restauração do serviço. Esse valor determina uma duração aceitável pela qual o serviço está comprometido.
- Objetivo de ponto de recuperação (RPO) — RPO é o tempo máximo aceitável desde o último ponto de recuperação de dados. Isso determina o que é considerado uma perda de dados aceitável entre o ponto de recuperação mais recente e a interrupção do serviço.

Assim como definir metas de disponibilidade, o RTO e o RPO também devem ser definidos no nível da carga de trabalho. Uma continuidade mais agressiva das operações ou a alta disponibilidade exigem maior investimento. Dito isso, nem todo aplicativo pode exigir ou exigir o mesmo nível de resiliência. Alinhar os proprietários de negócios e de TI para avaliar a importância dos aplicativos com base no impacto nos negócios e, em seguida, classificá-los adequadamente pode ajudar a fornecer um ponto de partida. As tabelas a seguir fornecem exemplos de hierarquização.

Esta tabela mostra um exemplo de escalonamento de resiliência para contratos de nível de serviço (). SLAs

Nível de resiliência	Disponibilidade do SLA	Tempo de inatividade aceitável/ano
Platina	99,99%	52.60 minutos
Ouro	99,90%	8.7 horas
Prata	99,5%	1,83 dias

A tabela a seguir mostra um exemplo de escalonamento de resiliência para RTO e RPO.

Nível de resiliência	RTO máximo	RPO máximo	Critérios	Custo
Platina	15 minutos	5 minutos	Cargas de trabalho de missão crítica	$$$
Ouro	15 minutos — 6 horas	2 horas	Cargas de trabalho importantes, mas não essenciais	$$
Prata	6 horas — alguns dias	24 horas	Cargas de trabalho não críticas	$

Ao projetar cargas de trabalho para resiliência, considere a relação entre alta disponibilidade e continuidade das operações. Por exemplo, se uma carga de trabalho exigir disponibilidade de 99,99%, não mais do que 53 minutos de inatividade por ano são toleráveis. Pode levar pelo menos 5 minutos para detectar uma falha e outros 10 minutos para que um operador se envolva, tome decisões sobre as etapas de recuperação e execute essas etapas. Não é incomum levar de 30 a 45 minutos para se recuperar de um único problema. Nesse caso, é vantajoso ter uma estratégia multirregional para fornecer uma instância isolada que elimine o impacto correlacionado. Isso permite operações contínuas por meio de falhas dentro de um tempo limitado, enquanto você faz a triagem

da deficiência inicial de forma independente. É aqui que é necessário definir o tempo de recuperação limitado apropriado e garantir o alinhamento.

Uma abordagem multirregional pode ser apropriada para cargas de trabalho de missão crítica que têm necessidades extremas de disponibilidade (por exemplo, disponibilidade de 99,99 por cento ou mais) ou requisitos rigorosos de continuidade de operações que só podem ser atendidos com o failover em outra região. No entanto, esses requisitos geralmente são aplicáveis somente a um pequeno subconjunto do portfólio de carga de trabalho de uma empresa que tem um tempo de recuperação limitado medido em minutos ou horas. A menos que um aplicativo precise de um tempo de recuperação de minutos ou algumas horas, talvez seja melhor esperar que uma interrupção regional do aplicativo seja corrigida na região afetada. Essa abordagem geralmente está alinhada com cargas de trabalho de nível inferior.

Antes de implementar uma arquitetura multirregional, os tomadores de decisão de negócios e as equipes técnicas devem estar alinhados sobre as implicações de custo, incluindo fatores de custo operacional e de infraestrutura. Uma arquitetura típica de várias regiões pode ter um custo duas vezes maior do que uma abordagem de região única. Embora existam vários padrões multirregionais para a continuidade dos negócios, como operar com espera quente, espera quente ou luz piloto, o padrão com o menor risco de atingir os objetivos de recuperação envolverá a execução de espera dinâmica e dobrará o custo de sua carga de trabalho.

Orientação-chave

- As metas de disponibilidade e continuidade das operações, como RTO e RPO, devem ser estabelecidas por carga de trabalho e alinhadas com os negócios e as partes interessadas de TI.
- A maioria das metas de disponibilidade e continuidade das operações pode ser alcançada em uma única região. Para metas que não podem ser alcançadas em uma única região, considere a opção Multirregião com uma visão clara das compensações entre custo, complexidade e benefícios.

Fundamental multirregional 2: compreender os dados

Gerenciar dados não é um problema trivial quando você adota arquiteturas multirregionais. A distância geográfica entre regiões impõe uma latência inevitável que se manifesta como o tempo necessário para replicar dados entre regiões. Serão necessárias compensações entre disponibilidade, consistência de dados e introdução de maior latência em uma carga de trabalho que usa uma arquitetura multirregional. Se você usa replicação assíncrona ou síncrona, precisará modificar seu aplicativo para lidar com as mudanças comportamentais impostas pela tecnologia de replicação. Os desafios relacionados à consistência e latência dos dados tornam muito difícil transformar um aplicativo existente projetado para uma única região em várias regiões. Compreender os requisitos de consistência de dados e os padrões de acesso aos dados para cargas de trabalho específicas é fundamental para ponderar as compensações.

2.a: Entendendo os requisitos de consistência de dados

O teorema CAP fornece uma referência para raciocinar sobre as compensações entre consistência de dados, disponibilidade e partições de rede. Somente dois desses requisitos podem ser satisfeitos ao mesmo tempo para uma carga de trabalho. Por definição, uma arquitetura multirregional inclui partições de rede entre regiões, então você precisa escolher entre disponibilidade e consistência.

Se você selecionar a disponibilidade de dados entre regiões, não incorrerá em latência significativa durante as operações de gravação transacional, pois a dependência da replicação assíncrona de dados comprometidos entre regiões resulta em redução da consistência entre regiões até que a replicação seja concluída. Com a replicação assíncrona, quando há uma falha na região primária, há uma grande probabilidade de que as operações de gravação estejam pendentes de replicação da região primária. Isso leva a um cenário em que os dados mais recentes ficam indisponíveis até que a replicação seja retomada, e um processo de reconciliação é necessário para lidar com transações em andamento que não foram replicadas da região que sofreu a interrupção. Esse cenário exige entender sua lógica de negócios e criar um processo específico para reproduzir a transação ou comparar os armazenamentos de dados entre regiões.

Para cargas de trabalho em que a replicação assíncrona é preferida, você pode usar serviços como Amazon Aurora e Amazon DynamoDB para replicação assíncrona entre regiões. Tanto os bancos de dados globais do Amazon Aurora quanto as tabelas globais do Amazon DynamoDB têm métricas padrão da CloudWatchAmazon para ajudar a monitorar o atraso na replicação. Um banco de dados global do Aurora consiste em uma região primária em que seus dados são gravados e em até cinco

regiões secundárias somente para leitura. As tabelas globais do DynamoDB consistem em tabelas de réplica multiativas em qualquer número de regiões nas quais seus dados são gravados e lidos.

Projetar a carga de trabalho para aproveitar as arquiteturas orientadas por eventos é um benefício para uma estratégia multirregional, pois significa que a carga de trabalho pode adotar a replicação assíncrona de dados e permitir a reconstrução do estado por meio da repetição de eventos. Como os serviços de streaming e mensagens armazenam em buffer os dados da carga útil das mensagens em uma única região, um processo regional de failover ou failback deve incluir um mecanismo para redirecionar os fluxos de dados de entrada do cliente. O processo também deve reconciliar cargas úteis em voo ou não entregues armazenadas na região que sofreu a interrupção.

Se você escolher o requisito de consistência do CAP e usar um banco de dados replicado de forma síncrona em todas as regiões para oferecer suporte aos aplicativos que são executados simultaneamente em várias regiões, você remove o risco de perda de dados e mantém os dados sincronizados entre as regiões. No entanto, isso introduz características de latência mais altas, porque as gravações precisam se comprometer com mais de uma região, e as regiões podem estar a centenas ou milhares de quilômetros umas das outras. Você precisa considerar essa característica de latência no design do seu aplicativo. Além disso, a replicação síncrona pode introduzir a chance de falhas correlacionadas, pois as gravações precisarão ser comprometidas em mais de uma região para serem bem-sucedidas. Se houver uma deficiência em uma região, você precisará formar um quórum para que as gravações sejam bem-sucedidas. Isso normalmente envolve configurar seu banco de dados em três regiões e estabelecer um quórum de duas em cada três regiões. Tecnologias como a Paxos podem ajudar a replicar e confirmar dados de forma síncrona, mas exigem um investimento significativo do desenvolvedor.

Quando as gravações envolvem replicação síncrona em várias regiões para atender aos fortes requisitos de consistência, a latência de gravação aumenta em uma ordem de magnitude. Uma latência de gravação mais alta não é algo que você normalmente possa adaptar em um aplicativo sem alterações significativas, como revisitar o tempo limite e a estratégia de repetição do seu aplicativo. Idealmente, isso deve ser levado em consideração quando o aplicativo está sendo projetado pela primeira vez. Para cargas de trabalho multirregionais em que a replicação síncrona é uma prioridade,AWS Partner as soluções podem ajudar.

2.b: Entendendo os padrões de acesso aos dados

Os padrões de acesso aos dados da carga de trabalho exigem muita leitura ou gravação. Compreender essa característica para uma carga de trabalho específica ajudará você a selecionar uma arquitetura multirregional apropriada.

Para cargas de trabalho com uso intenso de leitura, como conteúdo estático totalmente somente para leitura, você pode obter uma arquitetura multirregional ativa-ativa que tenha menos complexidade de engenharia quando comparada a uma carga de trabalho com muita gravação. Servir conteúdo estático na borda usando uma rede de entrega de conteúdo (CDN) garante a disponibilidade ao armazenar em cache o conteúdo mais próximo do usuário final. Usar conjuntos de recursos como failover de origem na Amazon CloudFront pode ajudar a conseguir isso. Outra opção é implantar a computação sem estado em várias regiões e usar o DNS para direcionar os usuários para a região mais próxima para ler o conteúdo. Você pode usar o Amazon Route 53 com uma política de roteamento de geolocalização para conseguir isso.

Para cargas de trabalho de leitura intensiva que têm uma porcentagem maior de tráfego de leitura do que tráfego de gravação, você pode usar uma estratégia global de leitura local e gravação. Isso significa que todas as solicitações de gravação vão para um banco de dados em uma região específica, os dados são replicados de forma assíncrona para todas as outras regiões e as leituras podem ser feitas em qualquer região. Essa abordagem exige que uma carga de trabalho adote uma consistência eventual, porque as leituras locais podem se tornar obsoletas como resultado do aumento da latência na replicação de gravações entre regiões.

Os bancos de dados globais do Aurora podem ajudar a provisionar réplicas de leitura em uma região de espera que pode lidar exclusivamente com todo o tráfego de leitura localmente e provisionar um único armazenamento de dados primário em uma região específica para lidar com o tráfego de gravação. Os dados são replicados de forma assíncrona do banco de dados principal para bancos de dados stand-by (réplicas de leitura), e os bancos de dados stand-by podem ser promovidos a primários se você precisar fazer failover de operações para a região stand-by. Você também pode usar o DynamoDB nessa abordagem. As tabelas globais do DynamoDB podem provisionar tabelas de réplica em todas as regiões, cada uma delas escalável para suportar qualquer volume de tráfego local de leitura ou gravação. Quando uma aplicação grava dados em uma tabela-réplica em uma região, o DynamoDB propaga automaticamente a gravação para as outras tabelas-réplica nas demais regiões da . Com essa configuração, os dados são replicados de forma assíncrona de uma região primária definida para tabelas de réplica em regiões em espera. As tabelas de réplica em qualquer região sempre podem aceitar gravações, portanto, a promoção de uma região de espera para primária é gerenciada no nível do aplicativo. Novamente, a carga de trabalho precisa adotar uma consistência eventual, o que pode exigir que ela seja reescrita se não tiver sido projetada para isso desde o início.

Para cargas de trabalho com muita gravação, uma região primária deve ser selecionada e a capacidade de fazer failover para uma região em espera deve ser incorporada à carga de trabalho. Em comparação com uma abordagem ativo-ativa, uma abordagem de espera primária tem

vantagens adicionais. Isso ocorre porque, para uma arquitetura ativa-ativa, a carga de trabalho precisa ser reescrita para lidar com o roteamento inteligente para regiões, estabelecer afinidade de sessão, garantir transações idempotentes e lidar com possíveis conflitos.

A maioria das cargas de trabalho que usam uma abordagem multirregional para resiliência não exigirá uma abordagem ativo-ativa. Você pode usar uma estratégia de fragmentação para fornecer maior resiliência, limitando o escopo do impacto de uma deficiência em toda a base de clientes. Se você puder fragmentar efetivamente uma base de clientes, poderá selecionar diferentes regiões primárias para cada fragmento. Por exemplo, você pode fragmentar clientes para que metade dos clientes estejam alinhados à Região um e metade estejam alinhados à Região dois. Ao tratar as regiões como células, você pode criar uma abordagem celular multirregional, o que resulta na redução do escopo do impacto em sua carga de trabalho. Para obter mais informações, consulte a apresentação do AWS re:Invent sobre essa abordagem.

Você pode combinar a abordagem de fragmentação com uma abordagem de espera primária para fornecer recursos de failover para os fragmentos. Você precisará criar um processo de failover testado na carga de trabalho e também um processo de reconciliação de dados, para garantir a consistência transacional dos armazenamentos de dados após o failover. Eles serão abordados com mais detalhes posteriormente neste guia.

Orientação-chave

- Há uma grande probabilidade de que as gravações pendentes para replicação não sejam confirmadas na região de espera quando houver uma falha. Os dados ficarão indisponíveis até que a replicação seja retomada (supondo a replicação assíncrona).
- Como parte do failover, será necessário um processo de reconciliação de dados para garantir que um estado transacionalmente consistente seja mantido para armazenamentos de dados que usam replicação assíncrona. Isso requer uma lógica comercial específica e não é algo que seja tratado pelo próprio armazenamento de dados.
- Quando for necessária uma consistência forte, as cargas de trabalho precisarão ser modificadas para tolerar a latência necessária de um armazenamento de dados que se replica de forma síncrona.

Princípio fundamental da multirregião 3: Entender suas dependências de carga de trabalho

Uma carga de trabalho específica pode ter várias dependências em uma região, como dependências Serviços da AWS usadas, internas, dependências de terceiros, dependências de rede, certificados, chaves, segredos e parâmetros. Para garantir a operação da carga de trabalho durante um cenário de falha, não deve haver dependências entre a região principal e a região de espera; cada uma deve ser capaz de operar independentemente da outra. Para conseguir isso, examine todas as dependências na carga de trabalho para garantir que elas estejam disponíveis em cada região. Isso é necessário porque uma falha na região principal não deve afetar a região de espera. Além disso, você deve entender como a carga de trabalho opera quando uma dependência está em um estado degradado ou completamente indisponível, para que você possa criar soluções para lidar com isso adequadamente.

3.a: Serviços da AWS

Ao projetar uma arquitetura multirregional, é importante entender o Serviços da AWS que será usado, os recursos multirregionais desses serviços e quais soluções você precisará criar para atingir as metas multirregionais. Por exemplo, o Amazon Aurora e o Amazon DynamoDB podem replicar dados de forma assíncrona para uma região em espera. Todas as AWS service (Serviço da AWS) dependências precisarão estar disponíveis em todas as regiões nas quais uma carga de trabalho será executada. Para confirmar se os serviços que você usa estão disponíveis nas regiões desejadas, consulte a lista Serviços da AWS por região.

3.b: Dependências internas e de terceiros

Certifique-se de que as dependências internas de cada carga de trabalho estejam disponíveis nas regiões nas quais elas operam. Por exemplo, se a carga de trabalho for composta por muitos microsserviços, identifique todos os microsserviços que compõem uma capacidade de negócios e verifique se todos esses microsserviços estão implantados em cada região na qual a carga de trabalho opera. Como alternativa, defina uma estratégia para lidar com microsserviços que se tornam indisponíveis.

Chamadas entre regiões entre microsserviços dentro de uma carga de trabalho não são recomendadas, e o isolamento regional deve ser mantido. Isso ocorre porque a criação de

dependências entre regiões aumenta o risco de falhas correlacionadas, o que compensa os benefícios de implementações regionais isoladas da carga de trabalho. As dependências locais também podem fazer parte da carga de trabalho, por isso é importante entender como as características dessas integrações poderiam mudar se a região principal mudasse. Por exemplo, se a região de espera estiver localizada mais longe do ambiente local, o aumento da latência poderá ter um impacto negativo.

Compreender as soluções de software como serviço (SaaS), os kits de desenvolvimento de software (SDKs) e outras dependências de produtos de terceiros e a capacidade de testar cenários em que essas dependências estejam degradadas ou indisponíveis fornecerá mais informações sobre como a cadeia de sistemas opera e se comporta em diferentes modos de falha. Essas dependências podem estar no código do seu aplicativo, como gerenciar segredos externamente usando AWS Secrets Manager, ou podem envolver uma solução de cofre de terceiros (como HashiCorp) ou sistemas de autenticação que dependem de logins federados. Centro de Identidade do AWS IAM

Ter redundância quando se trata de dependências pode aumentar a resiliência. Se uma solução SaaS ou uma dependência de terceiros usa a mesma carga primária Região da AWS da carga de trabalho, trabalhe com o fornecedor para determinar se a postura de resiliência corresponde aos requisitos da carga de trabalho.

Além disso, esteja ciente do destino compartilhado entre a carga de trabalho e suas dependências, como aplicativos de terceiros. Se as dependências não estiverem disponíveis em (ou de) uma região secundária após um failover, a carga de trabalho pode não se recuperar totalmente.

3.c: Mecanismo de failover

O DNS é comumente usado como um mecanismo de failover para transferir o tráfego da região primária para uma região em espera. Analise e examine criticamente todas as dependências que o mecanismo de failover assume. Por exemplo, se sua carga de trabalho usa o Amazon Route 53, entender que o plano de controle está hospedado em `us-east-1` significa que você está se tornando dependente do plano de controle nessa região específica. Isso não é recomendado como parte de um mecanismo de failover se a região principal também for `us-east-1` porque cria um único ponto de falha. Se você usar outro mecanismo de failover, deverá ter uma compreensão profunda dos cenários nos quais o failover não funcionaria conforme o esperado e, em seguida, planejar a contingência ou desenvolver um novo mecanismo, se necessário. O switch de região do Amazon Application Recovery Controller (ARC) é um serviço de recuperação multirregional totalmente gerenciado que você pode usar como mecanismo de failover.

Conforme discutido na seção anterior, todos os microsserviços que fazem parte de um recurso de negócios precisam estar disponíveis em cada região na qual a carga de trabalho é implantada. Como parte da estratégia de failover, todos os microsserviços que fazem parte da capacidade de negócios devem fazer o failover juntos para eliminar a chance de chamadas entre regiões. Como alternativa, se os microsserviços falharem de forma independente, existe a possibilidade de comportamentos indesejáveis, como microsserviços potencialmente fazendo chamadas entre regiões. Isso introduz latência e pode fazer com que a carga de trabalho fique indisponível durante o tempo limite do cliente.

3.d: Dependências de configuração

Certificados, chaves, segredos, Amazon Machine Images (AMIs), imagens de contêineres e parâmetros fazem parte da análise de dependência necessária ao projetar uma arquitetura multirregional. Sempre que possível, é melhor localizar esses componentes em cada região para que eles não tenham um destino compartilhado entre as regiões para essas dependências. Por exemplo, você deve variar as datas de expiração dos certificados para evitar um cenário em que um certificado expirado (com alarmes definidos para “notificar com antecedência”) afete várias regiões.

As chaves e segredos de criptografia também devem ser específicos da região. Dessa forma, se houver um erro na rotação de uma chave ou segredo, o impacto será limitado a uma região específica.

Por fim, todos os parâmetros da carga de trabalho devem ser armazenados localmente para que a carga de trabalho seja recuperada na região específica.

Orientação-chave

- Uma arquitetura multirregional se beneficia da separação física e lógica entre regiões. A introdução de dependências entre regiões na camada de aplicação elimina esse benefício. Evite essas dependências.
- Os controles de failover devem funcionar sem dependências na região principal.
- O failover deve ser coordenado em toda a jornada do usuário para eliminar a possibilidade de maior latência e dependência de chamadas entre regiões.

Fundamental multirregional 4: prontidão operacional

Operar uma carga de trabalho multirregional é uma tarefa complexa que vem com desafios operacionais específicos de uma arquitetura multirregional. Isso inclui Conta da AWS gerenciamento, processos de implantação reformulados, criação de uma estratégia de observabilidade em várias regiões, criação e teste de processos de recuperação e, em seguida, gerenciamento do custo. Uma análise de prontidão operacional (ORR) pode ajudar as equipes a preparar uma carga de trabalho para a produção, seja ela sendo executada em uma única região ou em várias regiões.

4.a: gestão Conta da AWS

Para implantar uma carga de trabalho em todas as regiões Regiões da AWS, certifique-se de que haja paridade em todas as AWS service (Serviço da AWS) cotas de uma conta em todas as regiões. Primeiro, identifique tudo o Serviços da AWS que faz parte da arquitetura, analise o uso planejado nas regiões de espera e, em seguida, compare o uso planejado com o uso atual. Em alguns casos, se a região em espera não tiver sido usada antes, você poderá consultar as cotas de serviço padrão para entender o ponto de partida. Em seguida, em todos os serviços que serão usados, solicite um aumento de cota usando o console Service Quotas (é necessário fazer login) ou. APIs

Configure funções AWS Identity and Access Management (IAM) em cada região para dar aos operadores, ferramentas de automação e Serviços da AWS as permissões apropriadas aos recursos dentro da região em espera. Para obter isolamento regional para arquiteturas multirregionais, isole as funções por região. Certifique-se de que as permissões estejam em vigor antes de ativar uma região em espera.

4.b: Práticas de implantação

Os recursos multirregionais podem complicar a implantação de uma carga de trabalho em várias regiões. Você precisa se certificar de implantar em uma região por vez. Por exemplo, se você usar uma abordagem ativa-passiva, deve implantar primeiro na região primária e depois na região de espera. AWS CloudFormationajuda você a implantar a infraestrutura em uma ou várias regiões e pode ser personalizada de acordo com suas necessidades. AWS CodePipelineajuda você a criar um pipeline de integration/continuous entrega contínua (CI/CD), que tem ações entre regiões que permitem a implantação em regiões diferentes da região em que o pipeline está. Isso, combinado com estratégias robustas de implantação, como azul/verde, permite uma implantação de tempo de inatividade mínimo a zero.

No entanto, a implantação de recursos de monitoramento de estado pode se tornar mais complexa quando o estado do aplicativo ou dos dados não é externalizado para um armazenamento persistente. Nessas situações, adapte cuidadosamente o processo de implantação para atender às suas necessidades. Projete o pipeline e o processo de implantação para implantar em uma região por vez, em vez de implantar em várias regiões simultaneamente. Isso reduz a chance de falhas correlacionadas entre as regiões. Para saber mais sobre as técnicas que a Amazon usa para automatizar implantações de software, consulte o artigo da AWS Builders' Library Automatizando implantações seguras e sem intervenção.

4.c: Observabilidade

Ao projetar para Multirregião, considere como você monitorará a integridade de todos os componentes em cada região para obter uma visão holística da saúde regional. Isso pode incluir métricas de monitoramento para atraso na replicação, o que não é considerado para uma carga de trabalho de uma única região.

Ao criar uma arquitetura multirregional, considere também observar o desempenho da carga de trabalho nas regiões em espera. Isso inclui fazer exames de saúde e canários (testes sintéticos) funcionando na região de espera para fornecer uma visão externa da saúde da região primária. Além disso, você pode usar o Amazon CloudWatch Internet Monitor para entender o estado da rede externa e o desempenho de suas cargas de trabalho do ponto de vista do usuário final. A região primária deve ter a mesma observabilidade para monitorar a região em espera.

Os canários da região de espera devem monitorar as métricas de experiência do cliente para determinar a integridade geral da carga de trabalho. Isso é necessário porque, se houver um problema na região primária, a observabilidade na primária pode ser prejudicada e afetaria sua capacidade de avaliar a saúde da carga de trabalho.

Nesse caso, observar fora dessa região pode fornecer uma visão. Essas métricas devem ser agrupadas em painéis disponíveis em cada região e alarmes criados em cada região. Por ser um serviço regional, CloudWatché necessário ter alarmes nas duas regiões. Esses dados de monitoramento serão usados para fazer a chamada para o failover de uma região primária para uma de espera.

4.d: Processos e procedimentos

O melhor momento para responder à pergunta: “Quando devo falhar?” é muito antes de você precisar. Defina planos de recuperação que incluam pessoas, processos e tecnologia bem antes

de um problema e teste-os regularmente. Decida sobre uma estrutura de decisão de recuperação. Se houver um processo de recuperação bem praticado e o tempo de recuperação for bem compreendido, você poderá optar por iniciar o processo de recuperação usando um failover que atenda à meta de RTO. Esse momento pode ocorrer imediatamente após a identificação de um problema com o aplicativo na região principal ou pode ser um evento posterior quando as opções de recuperação do aplicativo na região tiverem sido esgotadas.

A ação de failover em si deve ser 100% automatizada, mas a decisão de ativar o failover deve ser tomada por humanos, geralmente um pequeno número de indivíduos predeterminados na organização. Essas pessoas devem considerar a perda de dados e informações sobre o evento. Além disso, os critérios para um failover precisam ser claramente definidos e compreendidos globalmente dentro da organização. Para definir e concluir esses processos, você pode usar o switch de região do Amazon Application Recovery Controller (ARC), que permite a end-to-end automação completa e garante a consistência dos processos em execução durante o teste e o failover.

Quando você cria um plano de mudança de região, ele replica automaticamente seu plano nas regiões primária e de reserva para garantir que não haja dependência de uma única região. Quando essa automação estiver em vigor, defina e siga uma cadência de testes regular. Isso garante que, quando houver um evento real, a resposta siga um processo bem definido e praticado no qual a organização tenha confiança. Também é importante considerar as tolerâncias estabelecidas para os processos de reconciliação de dados. Confirme se o processo proposto atende aos RPO/RTO requisitos estabelecidos.

4.e: Teste

Ter uma abordagem de recuperação não testada é igual a não ter uma abordagem de recuperação. Um nível básico de teste seria executar um procedimento de recuperação para mudar a região operacional do seu aplicativo. Às vezes, isso é chamado de abordagem de rotação de aplicativos. Recomendamos que você crie a capacidade de mudar as regiões para sua postura operacional normal; no entanto, esse teste por si só não é suficiente.

O teste de resiliência também é fundamental para validar a abordagem de recuperação de um aplicativo. Isso envolve a injeção de cenários de falha específicos, a compreensão de como o aplicativo e o processo de recuperação reagem e, em seguida, a implementação de quaisquer mitigações necessárias caso o teste não ocorra conforme o planejado. Testar seu procedimento de recuperação na ausência de erros não lhe dirá como seu aplicativo se comporta como um todo quando ocorrem falhas. Você deve desenvolver um plano para testar sua recuperação em

relação aos cenários de falha esperados. AWS Fault Injection Servicefornece uma lista crescente de cenários para você começar.

Isso é especialmente importante para aplicativos de alta disponibilidade, nos quais testes rigorosos são necessários para garantir que as metas de continuidade de negócios sejam atingidas. Testar proativamente os recursos de recuperação reduz o risco de falhas na produção, o que aumenta a confiança de que o aplicativo pode atingir o tempo de recuperação limitado desejado. Os testes regulares também criam experiência operacional, o que permite que a equipe se recupere de forma rápida e confiável das interrupções quando elas ocorrerem. Exercitar o elemento humano, ou processo, de sua abordagem de recuperação é tão importante quanto os aspectos técnicos.

4.f: Custo e complexidade

As implicações de custo de uma arquitetura multirregional são impulsionadas pelo maior uso da infraestrutura, pela sobrecarga operacional e pelo tempo de recursos. Conforme mencionado anteriormente, o custo da infraestrutura em uma região em espera é semelhante ao custo da infraestrutura em uma região primária durante o pré-provisionamento, portanto, dobra seu custo total. Provisione a capacidade de forma que seja suficiente para as operações diárias, mas ainda reserve capacidade de buffer suficiente para tolerar picos na demanda. Em seguida, configure os mesmos limites em cada região.

Além disso, se você estiver adotando uma arquitetura ativa-ativa, talvez seja necessário fazer alterações no nível do aplicativo para executar seu aplicativo com êxito em uma arquitetura multirregional. Essas mudanças podem consumir muito tempo e recursos para projetar e operar. No mínimo, as organizações precisam gastar tempo entendendo as dependências técnicas e comerciais em cada região e projetando processos de failover e failback.

As equipes também devem realizar exercícios normais de failover e failback para se sentirem confortáveis com os runbooks que seriam usados durante um evento. Embora esses exercícios sejam cruciais para obter o resultado esperado de um investimento em várias regiões, eles representam um custo de oportunidade e consomem tempo e recursos de outras atividades.

4.g: Estratégia organizacional de failover multirregional

Regiões da AWS forneça limites de isolamento de falhas que evitem falhas correlacionadas e contenham o impacto das AWS service (Serviço da AWS) deficiências, quando elas ocorrem, em uma única região. Você pode usar esses limites de falha para criar aplicativos multirregionais que consistem em réplicas independentes e isoladas de falhas em cada região para limitar cenários de

destino compartilhado. Isso permite que você crie aplicativos multirregionais e use uma variedade de abordagens — do backup e da restauração ao piloto leve e ao ativo-ativo — para implementar sua arquitetura multirregional. No entanto, os aplicativos normalmente não operam isoladamente, portanto, considere os componentes que você usará e suas dependências como parte de sua estratégia de failover. Geralmente, vários aplicativos trabalham juntos para oferecer suporte a uma história de usuário, que é um recurso específico oferecido a um usuário final, como publicar uma foto e uma legenda em um aplicativo de mídia social ou fazer o check-out em um site de comércio eletrônico. Por isso, você deve desenvolver uma estratégia organizacional de failover multirregional que forneça a coordenação e a consistência necessárias para que sua abordagem seja bem-sucedida.

Há quatro estratégias de alto nível que as organizações podem escolher para orientar uma abordagem multirregional. Elas estão listadas da abordagem mais granular à mais ampla:

- Failover em nível de componente
- Failover de aplicativo individual
- Failover do gráfico de dependências
- Failover de todo o portfólio de aplicativos

Cada estratégia tem vantagens e desvantagens e aborda diferentes desafios, incluindo flexibilidade na tomada de decisões de failover, capacidade de testar as combinações de failover, presença de comportamento modal e investimento organizacional em planejamento e implementação. Para se aprofundar em cada estratégia com mais detalhes, consulte a postagem do AWS blog Criação de uma estratégia organizacional de failover multirregional.

Orientação-chave

- Analise todas as AWS service (Serviço da AWS) cotas para garantir que elas estejam em paridade em todas as regiões nas quais a carga de trabalho será operada.
- O processo de implantação deve ter como alvo uma região por vez, em vez de envolver várias regiões simultaneamente.
- Métricas adicionais, como atraso na replicação, são específicas para cenários multirregionais e devem ser monitoradas.
- Estenda o monitoramento da carga de trabalho além da região principal. Monitore as métricas de experiência do cliente para cada região e meça esses dados de fora de cada região na qual uma carga de trabalho está sendo executada.

- Teste o failover e o failback regularmente. Implemente um único runbook para processos de failover e failback e use-o tanto para testes quanto para eventos ao vivo. Os runbooks para testes e eventos ao vivo não devem ser diferentes.
- Entenda as vantagens e desvantagens das estratégias de failover. Implemente um gráfico de dependências ou uma estratégia completa do portfólio de aplicativos.

Conclusão e atributos

Este guia abordou casos de uso comuns de arquiteturas multirregionais, os fundamentos da implementação dessas arquiteturas e as implicações dessa abordagem. Você pode aplicar esses fundamentos a qualquer carga de trabalho e usar as informações como uma estrutura para ajudar a decidir se uma arquitetura multirregional é a abordagem certa para sua empresa.

Para obter mais informações, consulte os seguintes recursos:

- AWS Centro de Arquitetura
- AWS Well-Architected Framework
- AWS Well-Architected Tool
- Criação de uma estratégia organizacional de failover multirregional (AWS postagem no blog)
- AWS Capacidades multirregionais (artigo do AWS re:POST)

Histórico do documento

A tabela a seguir descreve alterações significativas feitas neste guia. Se desejar receber notificações sobre futuras atualizações, inscreva-se em um feed RSS.

Alteração	Descrição	Data
Atualizações para o switch ARC Region	Nas seções 3.c e 4.d, foram adicionadas informações sobre o switch de região do Amazon Application Recovery Controller (ARC) para lidar com tarefas de recuperação.	29 de setembro de 2025
Atualizações	Atualizações em todo o guia.	27 de dezembro de 2024
Publicação inicial	—	20 de dezembro de 2022

AWS Glossário de orientação prescritiva

A seguir estão os termos comumente usados em estratégias, guias e padrões fornecidos pela Orientação AWS Prescritiva. Para sugerir entradas, use o link Fornecer feedback no final do glossário.

Números

7 Rs

Sete estratégias comuns de migração para mover aplicações para a nuvem. Essas estratégias baseiam-se nos 5 Rs identificados pela Gartner em 2011 e consistem em:

- Refatorar/rearquitetar: mova uma aplicação e modifique sua arquitetura aproveitando ao máximo os recursos nativos de nuvem para melhorar a agilidade, a performance e a escalabilidade. Isso normalmente envolve a portabilidade do sistema operacional e do banco de dados. Exemplo: migrar seu banco de dados Oracle on-premises para o Amazon Aurora Edição Compatível com PostgreSQL.
- Redefinir a plataforma (mover e redefinir [mover e redefinir (lift-and-reshape]): mova uma aplicação para a nuvem e introduza algum nível de otimização a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Amazon Relational Database Service (Amazon RDS) para Oracle na Nuvem AWS.
- Recomprar (drop and shop): mude para um produto diferente, normalmente migrando de uma licença tradicional para um modelo SaaS. Exemplo: migrar seu sistema de gerenciamento de relacionamento com o cliente (CRM) para o Salesforce.com.
- Redefinir a hospedagem (mover sem alterações [lift-and-shift])mover uma aplicação para a nuvem sem fazer nenhuma alteração a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Oracle em uma instância do EC2 na Nuvem AWS.
- Realocar (mover o hipervisor sem alterações [hypervisor-level lift-and-shift]): mover a infraestrutura para a nuvem sem comprar novo hardware, reescrever aplicações ou modificar suas operações existentes. Você migra servidores de uma plataforma on-premises para um serviço de nuvem para a mesma plataforma. Exemplo: Migrar um Microsoft Hyper-V aplicativo para o. AWS
- Reter (revisitar): mantenha as aplicações em seu ambiente de origem. Isso pode incluir aplicações que exigem grande refatoração, e você deseja adiar esse trabalho para um

momento posterior, e aplicações antigas que você deseja manter porque não há justificativa comercial para migrá-las.

- Retirar: desative ou remova aplicações que não são mais necessárias em seu ambiente de origem.

A

ABAC

Consulte controle de acesso baseado em atributo.

serviços abstraídos

Veja serviços gerenciados.

ACID

Veja atomicidade, consistência, isolamento, durabilidade.

migração ativa-ativa

Um método de migração de banco de dados no qual os bancos de dados de origem e de destino são mantidos em sincronia (por meio de uma ferramenta de replicação bidirecional ou operações de gravação dupla), e ambos os bancos de dados lidam com transações de aplicações conectadas durante a migração. Esse método oferece suporte à migração em lotes pequenos e controlados, em vez de exigir uma substituição única. É mais flexível, mas exige mais trabalho do que a migração ativa-passiva.

migração ativa-passiva

Um método de migração de banco de dados em que os bancos de dados de origem e de destino são mantidos em sincronia, mas somente o banco de dados de origem manipula as transações das aplicações conectadas, enquanto os dados são replicados no banco de dados de destino. O banco de dados de destino não aceita nenhuma transação durante a migração.

AGGREGATE FUNCTION

Uma função SQL que opera em um grupo de linhas e calcula um único valor de retorno para o grupo. Exemplos de funções agregadas incluem `SUM` e `MAX`.

AI

Veja inteligência artificial.

AIOps

Veja operações de inteligência artificial.

anonimização

O processo de excluir permanentemente informações pessoais em um conjunto de dados. A anonimização pode ajudar a proteger a privacidade pessoal. Dados anônimos não são mais considerados dados pessoais.

antipadrões

Uma solução frequentemente usada para um problema recorrente em que a solução é contraproducente, ineficaz ou menos eficaz do que uma alternativa.

controle de aplicações

Uma abordagem de segurança que permite o uso somente de aplicações aprovadas para ajudar a proteger um sistema contra malware.

portfólio de aplicações

Uma coleção de informações detalhadas sobre cada aplicação usada por uma organização, incluindo o custo para criar e manter a aplicação e seu valor comercial. Essas informações são fundamentais para o processo de descoberta e análise de portfólio e ajudam a identificar e priorizar as aplicações a serem migradas, modernizadas e otimizadas.

inteligência artificial (IA)

O campo da ciência da computação que se dedica ao uso de tecnologias de computação para desempenhar funções cognitivas normalmente associadas aos humanos, como aprender, resolver problemas e reconhecer padrões. Para obter mais informações, consulte O que é inteligência artificial?

operações de inteligência artificial (AIOps)

O processo de usar técnicas de machine learning para resolver problemas operacionais, reduzir incidentes operacionais e intervenção humana e aumentar a qualidade do serviço. Para obter mais informações sobre como AIOps é usado na estratégia de AWS migração, consulte o guia de integração de operações.

criptografia assimétrica

Um algoritmo de criptografia que usa um par de chaves, uma chave pública para criptografia e uma chave privada para descriptografia. É possível compartilhar a chave pública porque ela não é usada na descriptografia, mas o acesso à chave privada deve ser altamente restrito.

atomicidade, consistência, isolamento, durabilidade (ACID)

Um conjunto de propriedades de software que garantem a validade dos dados e a confiabilidade operacional de um banco de dados, mesmo no caso de erros, falhas de energia ou outros problemas.

controle de acesso por atributo (ABAC)

A prática de criar permissões minuciosas com base nos atributos do usuário, como departamento, cargo e nome da equipe. Para obter mais informações, consulte ABAC AWS na documentação AWS Identity and Access Management (IAM).

fonte de dados autorizada

Um local onde você armazena a versão principal dos dados, que é considerada a fonte de informações mais confiável. Você pode copiar dados da fonte de dados autorizada para outros locais com o objetivo de processar ou modificar os dados, como anonimizá-los, redigi-los ou pseudonimizá-los.

Zona de disponibilidade

Um local distinto dentro de um Região da AWS que está isolado de falhas em outras zonas de disponibilidade e fornece conectividade de rede barata e de baixa latência a outras zonas de disponibilidade na mesma região.

AWS Estrutura de adoção da nuvem (AWS CAF)

Uma estrutura de diretrizes e melhores práticas AWS para ajudar as organizações a desenvolver um plano eficiente e eficaz para migrar com sucesso para a nuvem. AWS O CAF organiza a orientação em seis áreas de foco chamadas perspectivas: negócios, pessoas, governança, plataforma, segurança e operações. As perspectivas de negócios, pessoas e governança têm como foco habilidades e processos de negócios; as perspectivas de plataforma, segurança e operações concentram-se em habilidades e processos técnicos. Por exemplo, a perspectiva das pessoas tem como alvo as partes interessadas que lidam com recursos humanos (RH), funções de pessoal e gerenciamento de pessoal. Nessa perspectiva, o AWS CAF fornece orientação para desenvolvimento, treinamento e comunicação de pessoas para ajudar a preparar a organização para a adoção bem-sucedida da nuvem. Para obter mais informações, consulte o site da AWS CAF e o whitepaper da AWS CAF.

AWS Estrutura de qualificação da carga de trabalho (AWS WQF)

Uma ferramenta que avalia as cargas de trabalho de migração do banco de dados, recomenda estratégias de migração e fornece estimativas de trabalho. AWS O WQF está incluído com AWS

Schema Conversion Tool ()AWS SCT. Ela analisa esquemas de banco de dados e objetos de código, código de aplicações, dependências e características de performance, além de fornecer relatórios de avaliação.

B

bot malicioso

Um bot destinado a causar disrupção ou danos a indivíduos ou organizações.

BCP

Veja planejamento de continuidade de negócios

gráfico de comportamento

Uma visualização unificada e interativa do comportamento e das interações de recursos ao longo do tempo. É possível usar um gráfico de comportamento com o Amazon Detective para examinar tentativas de login malsucedidas, chamadas de API suspeitas e ações similares. Para obter mais informações, consulte Dados em um gráfico de comportamento na documentação do Detective.

sistema big-endian

Um sistema que armazena o byte mais significativo antes. Veja também endianness.

classificação binária

Um processo que prevê um resultado binário (uma de duas classes possíveis). Por exemplo, seu modelo de ML pode precisar prever problemas como “Este e-mail é ou não é spam?” ou “Este produto é um livro ou um carro?”

filtro de bloom

Uma estrutura de dados probabilística e eficiente em termos de memória que é usada para testar se um elemento é membro de um conjunto.

blue/green deployment (implantação azul/verde)

Uma estratégia de implantação em que você cria dois ambientes separados, mas idênticos. Você executa a versão atual da aplicação em um ambiente (azul) e a nova versão da aplicação no outro ambiente (verde). Essa estratégia ajuda você a reverter rapidamente com o mínimo de impacto.

bot

Uma aplicação de software que executa tarefas automatizadas na internet e simula a atividade ou interação humana. Alguns bots são úteis ou benéficos, como crawlers da web que indexam informações na internet. Outros bots, conhecidos como bots maliciosos, têm como objetivo causar disrupção ou danos a indivíduos ou organizações.

botnet

Redes de bots infectadas por malware e sob o controle de uma única parte, conhecidas como bot herder ou operador de bots. Os botnets são o mecanismo mais conhecido para escalar bots e seu impacto.

ramo

Uma área contida de um repositório de código. A primeira ramificação criada em um repositório é a ramificação principal. Você pode criar uma nova ramificação a partir de uma ramificação existente e, em seguida, desenvolver recursos ou corrigir bugs na nova ramificação. Uma ramificação que você cria para gerar um recurso é comumente chamada de ramificação de recurso. Quando o recurso estiver pronto para lançamento, você mesclará a ramificação do recurso de volta com a ramificação principal. Para obter mais informações, consulte Sobre filiais (GitHub documentação).

Acesso de emergência

Em circunstâncias excepcionais e por meio de um processo aprovado, um meio rápido para um usuário obter acesso a um Conta da AWS que ele normalmente não tem permissão para acessar. Para obter mais informações, consulte o indicador Implement break-glass procedures nas orientações do AWS Well-Architected.

estratégia brownfield

A infraestrutura existente em seu ambiente. Ao adotar uma estratégia brownfield para uma arquitetura de sistema, você desenvolve a arquitetura de acordo com as restrições dos sistemas e da infraestrutura atuais. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

cache do buffer

A área da memória em que os dados acessados com mais frequência são armazenados.

capacidade de negócios

O que uma empresa faz para gerar valor (por exemplo, vendas, atendimento ao cliente ou marketing). As arquiteturas de microsserviços e as decisões de desenvolvimento podem

ser orientadas por recursos de negócios. Para obter mais informações, consulte a seção Organizados de acordo com as capacidades de negócios do whitepaper Executar microsserviços conteinerizados na AWS.

planejamento de continuidade de negócios (BCP)

Um plano que aborda o impacto potencial de um evento disruptivo, como uma migração em grande escala, nas operações e permite que uma empresa retome as operações rapidamente.

C

CAF

Veja AWS Cloud Adoption Framework.

implantação canário

O lançamento lento e incremental de uma versão para usuários finais. Quando estiver confiante, você implanta a nova versão e substitui a versão atual por completo.

CCoE

Veja Centro de Excelência da Nuvem.

CDC

Veja captura de dados de alteração.

captura de dados de alterações (CDC)

O processo de rastrear alterações em uma fonte de dados, como uma tabela de banco de dados, e registrar metadados sobre a alteração. É possível usar o CDC para várias finalidades, como auditar ou replicar alterações em um sistema de destino para manter a sincronização.

engenharia do caos

Introduzir intencionalmente falhas ou eventos disruptivos para testar a resiliência de um sistema. Você pode usar AWS Fault Injection Service (AWS FIS) para realizar experimentos que estressam suas AWS cargas de trabalho e avaliar sua resposta.

CI/CD

Veja integração e entrega contínuas.

classificação

Um processo de categorização que ajuda a gerar previsões. Os modelos de ML para problemas de classificação predizem um valor discreto. Os valores discretos são sempre diferentes uns dos outros. Por exemplo, um modelo pode precisar avaliar se há ou não um carro em uma imagem.

criptografia no lado do cliente

Criptografia de dados localmente, antes que o alvo os AWS service (Serviço da AWS) receba.

Centro de excelência em nuvem (CCoE)

Uma equipe multidisciplinar que impulsiona os esforços de adoção da nuvem em toda a organização, incluindo o desenvolvimento de práticas recomendas de nuvem, a mobilização de recursos, o estabelecimento de cronogramas de migração e a liderança da organização em transformações em grande escala. Para obter mais informações, consulte as publicações CCo E no blog de estratégia Nuvem AWS corporativa.

computação em nuvem

A tecnologia de nuvem normalmente usada para armazenamento de dados remoto e gerenciamento de dispositivos de IoT. A computação em nuvem é normalmente conectada à tecnologia de computação de borda.

modelo operacional em nuvem

Em uma organização de TI, o modelo operacional usado para criar, amadurecer e otimizar um ou mais ambientes de nuvem. Para obter mais informações, consulte Criar seu modelo operacional de nuvem.

estágios de adoção da nuvem

As quatro fases pelas quais as organizações normalmente passam ao migrar para a Nuvem AWS:

- Projeto: executar alguns projetos relacionados à nuvem para fins de prova de conceito e aprendizado
- Fundação — Fazer investimentos fundamentais para escalar sua adoção da nuvem (por exemplo, criar uma landing zone, definir um CCo E, estabelecer um modelo de operações)
- Migração: migrar aplicações individuais
- Reinvenção: otimizar produtos e serviços e inovar na nuvem

Esses estágios foram definidos por Stephen Orban na postagem do blog The Journey Toward Cloud-First & the Stages of Adoption no blog de estratégia Nuvem AWS empresarial. Para obter

informações sobre como eles se relacionam com a estratégia de AWS migração, consulte o guia de preparação para migração.

CMDB

Veja banco de dados de gerenciamento de configuração.

repositório de código

Um local onde o código-fonte e outros ativos, como documentação, amostras e scripts, são armazenados e atualizados por meio de processos de controle de versão. Os repositórios de nuvem comuns incluem o GitHub ou o Bitbucket Cloud. Cada versão do código é chamada de ramificação. Em uma estrutura de microsserviços, cada repositório é dedicado a uma única peça de funcionalidade. Um único pipeline de CI/CD pode usar vários repositórios.

cache frio

Um cache de buffer que está vazio, não está bem preenchido ou contém dados obsoletos ou irrelevantes. Isso afeta a performance porque a instância do banco de dados deve ler da memória principal ou do disco, um processo que é mais lento do que a leitura do cache do buffer.

dados frios

Dados que raramente são acessados e geralmente são históricos. Ao consultar esse tipo de dados, consultas lentas geralmente são aceitáveis. Mover esses dados para níveis ou classes de armazenamento de baixo desempenho e menos caros pode reduzir os custos.

visão computacional (CV)

Um campo de IA que usa machine learning para analisar e extrair informações de formatos visuais, como vídeos e imagens digitais. Por exemplo, a Amazon SageMaker AI fornece algoritmos de processamento de imagem para CV.

desvio de configuração

Em uma workload, uma alteração de configuração em relação ao estado esperado. Isso pode fazer com que a workload se torne incompatível e, normalmente, é gradual e não intencional.

banco de dados de gerenciamento de configuração (CMDB)

Um repositório que armazena e gerencia informações sobre um banco de dados e seu ambiente de TI, incluindo componentes de hardware e software e suas configurações. Normalmente, os dados de um CMDB são usados no estágio de descoberta e análise do portfólio da migração.

pacote de conformidade

Um conjunto de AWS Config regras e ações de remediação que você pode montar para personalizar suas verificações de conformidade e segurança. Você pode implantar um pacote de conformidade como uma entidade única em uma Conta da AWS região ou em uma organização usando um modelo YAML. Para obter mais informações, consulte Pacotes de conformidade na documentação. AWS Config

integração contínua e entrega contínua (CI/CD)

O processo de automatizar os estágios de origem, criação, teste, preparação e produção do processo de lançamento do software. CI/CD é comumente descrito como um pipeline. CI/CD pode ajudá-lo a automatizar processos, melhorar a produtividade, melhorar a qualidade do código e entregar com mais rapidez. Para obter mais informações, consulte Benefícios da entrega contínua. CD também pode significar implantação contínua. Para obter mais informações, consulte Entrega contínua versus implantação contínua.

CV

Veja visão computacional.

D

dados em repouso

Dados estacionários em sua rede, por exemplo, dados que estão em um armazenamento.

classificação de dados

Um processo para identificar e categorizar os dados em sua rede com base em criticalidade e confidencialidade. É um componente crítico de qualquer estratégia de gerenciamento de riscos de segurança cibernética, pois ajuda a determinar os controles adequados de proteção e retenção para os dados. A classificação de dados é um componente do pilar de segurança no AWS Well-Architected Framework. Para obter mais informações, consulte Classificação de dados.

desvio de dados

Uma variação significativa entre os dados de produção e os dados usados para treinar um modelo de ML ou uma alteração significativa nos dados de entrada ao longo do tempo. O desvio de dados pode reduzir a qualidade geral, a precisão e a imparcialidade das previsões do modelo de ML.

dados em trânsito

Dados que estão se movendo ativamente pela sua rede, como entre os recursos da rede.

data mesh

Um framework de arquitetura que fornece propriedade de dados distribuída e descentralizada com gerenciamento e governança centralizados.

minimização de dados

O princípio de coletar e processar apenas os dados estritamente necessários. Praticar a minimização de dados no Nuvem AWS pode reduzir os riscos de privacidade, os custos e a pegada de carbono de sua análise.

perímetro de dados

Um conjunto de proteções preventivas em seu AWS ambiente que ajudam a garantir que somente identidades confiáveis acessem recursos confiáveis das redes esperadas. Para obter mais informações, consulte Construindo um perímetro de dados em. AWS

pré-processamento de dados

A transformação de dados brutos em um formato que seja facilmente analisado por seu modelo de ML. O pré-processamento de dados pode significar a remoção de determinadas colunas ou linhas e o tratamento de valores ausentes, inconsistentes ou duplicados.

proveniência dos dados

O processo de rastrear a origem e o histórico dos dados ao longo de seu ciclo de vida, por exemplo, como os dados foram gerados, transmitidos e armazenados.

titular dos dados

Um indivíduo cujos dados estão sendo coletados e processados.

data warehouse

Um sistema de gerenciamento de dados compatível com business intelligence, como analytics. Os data warehouses geralmente contêm grandes quantidades de dados históricos e geralmente são usados para consultas e análises.

linguagem de definição de dados (DDL)

Instruções ou comandos para criar ou modificar a estrutura de tabelas e objetos em um banco de dados.

linguagem de manipulação de dados (DML)

Instruções ou comandos para modificar (inserir, atualizar e excluir) informações em um banco de dados.

DDL

Veja linguagem de definição de banco de dados.

deep ensemble

A combinação de vários modelos de aprendizado profundo para gerar previsões. Os deep ensembles podem ser usados para produzir uma previsão mais precisa ou para estimar a incerteza nas previsões.

Aprendizado profundo

Um subcampo do ML que usa várias camadas de redes neurais artificiais para identificar o mapeamento entre os dados de entrada e as variáveis-alvo de interesse.

defense-in-depth

Uma abordagem de segurança da informação na qual uma série de mecanismos e controles de segurança são cuidadosamente distribuídos por toda a rede de computadores para proteger a confidencialidade, a integridade e a disponibilidade da rede e dos dados nela contidos. Ao adotar essa estratégia AWS, você adiciona vários controles em diferentes camadas da AWS Organizations estrutura para ajudar a proteger os recursos. Por exemplo, uma defense-in-depth abordagem pode combinar autenticação multifatorial, segmentação de rede e criptografia.

administrador delegado

Em AWS Organizations, um serviço compatível pode registrar uma conta de AWS membro para administrar as contas da organização e gerenciar as permissões desse serviço. Essa conta é chamada de administrador delegado para esse serviço Para obter mais informações e uma lista de serviços compatíveis, consulte Serviços que funcionam com o AWS Organizations na documentação do AWS Organizations .

implantação

O processo de criar uma aplicação, novos recursos ou correções de código disponíveis no ambiente de destino. A implantação envolve a implementação de mudanças em uma base de código e, em seguida, a criação e execução dessa base de código nos ambientes da aplicação

ambiente de desenvolvimento

Veja ambiente.

controle detectivo

Um controle de segurança projetado para detectar, registrar e alertar após a ocorrência de um evento. Esses controles são uma segunda linha de defesa, alertando você sobre eventos de segurança que contornaram os controles preventivos em vigor. Para obter mais informações, consulte Controles detectivos em Como implementar controles de segurança na AWS.

mapeamento do fluxo de valor de desenvolvimento (DVSM)

Um processo usado para identificar e priorizar restrições que afetam negativamente a velocidade e a qualidade em um ciclo de vida de desenvolvimento de software. O DVSM estende o processo de mapeamento do fluxo de valor originalmente projetado para práticas de manufatura enxuta. Ele se concentra nas etapas e equipes necessárias para criar e movimentar valor por meio do processo de desenvolvimento de software.

gêmeo digital

Uma representação virtual de um sistema real, como um prédio, fábrica, equipamento industrial ou linha de produção. Os gêmeos digitais oferecem suporte à manutenção preditiva, ao monitoramento remoto e à otimização da produção.

tabela de dimensões

Em um esquema em estrela, uma tabela menor que contém atributos de dados sobre dados quantitativos em uma tabela de fatos. Os atributos da tabela de dimensões geralmente são campos de texto ou números discretos que se comportam como texto. Esses atributos normalmente são usados para restringir consultas, filtrar e rotular conjuntos de resultados.

desastre

Um evento que impede que uma workload ou sistema cumpra seus objetivos de negócios em seu local principal de implantação. Esses eventos podem ser desastres naturais, falhas técnicas ou o resultado de ações humanas, como configuração incorreta não intencional ou ataque de malware.

Recuperação de desastres (RD)

A estratégia e o processo que você usa para minimizar o tempo de inatividade e a perda de dados causados por um disastre. Para obter mais informações, consulte Recuperação de desastres de cargas de trabalho em AWS: Recuperação na nuvem no AWS Well-Architected Framework.

DML

Veja linguagem de manipulação de banco de dados.

design orientado por domínio

Uma abordagem ao desenvolvimento de um sistema de software complexo conectando seus componentes aos domínios em evolução, ou principais metas de negócios, atendidos por cada componente. Esse conceito foi introduzido por Eric Evans em seu livro, Design orientado por domínio: lidando com a complexidade no coração do software (Boston: Addison-Wesley Professional, 2003). Para obter informações sobre como usar o design orientado por domínio com o padrão strangler fig, consulte Modernizar incrementalmente os serviços web herdados do Microsoft ASP.NET (ASMX) usando contêineres e o Amazon API Gateway.

DR

Veja recuperação de desastres.

Detecção da oscilação

Rastreamento de desvios de uma configuração de linha de base. Por exemplo, você pode usar AWS CloudFormation para detectar desvios nos recursos do sistema ou AWS Control Tower para detectar mudanças em seu landing zone que possam afetar a conformidade com os requisitos de governança.

DVSM

Veja mapeamento do fluxo de valor de desenvolvimento.

E

EDA

Veja análise exploratória de dados.

EDI

Veja intercâmbio eletrônico de dados.

computação de borda

A tecnologia que aumenta o poder computacional de dispositivos inteligentes nas bordas de uma rede de IoT. Quando comparada com a computação em nuvem, a computação de borda pode reduzir a latência da comunicação e melhorar o tempo de resposta.

intercâmbio eletrônico de dados (EDI)

A troca automatizada de documentos comerciais entre organizações. Para obter mais informações, consulte O que é EDI (Intercâmbio eletrônico de dados)?.

criptografia

Um processo de computação que transforma dados de texto simples, legíveis por humanos, em texto cifrado.

chave de criptografia

Uma sequência criptográfica de bits aleatórios que é gerada por um algoritmo de criptografia. As chaves podem variar em tamanho, e cada chave foi projetada para ser imprevisível e exclusiva.

endianismo

A ordem na qual os bytes são armazenados na memória do computador. Os sistemas big-endian armazenam o byte mais significativo antes. Os sistemas little-endian armazenam o byte menos significativo antes.

endpoint

Veja endpoint de serviço.

serviço de endpoint

Um serviço que pode ser hospedado em uma nuvem privada virtual (VPC) para ser compartilhado com outros usuários. Você pode criar um serviço de endpoint com AWS PrivateLink e conceder permissões a outros diretores Contas da AWS ou a AWS Identity and Access Management (IAM). Essas contas ou entidades principais podem se conectar ao serviço de endpoint de maneira privada criando endpoints da VPC de interface. Para obter mais informações, consulte Criar um serviço de endpoint na documentação do Amazon Virtual Private Cloud (Amazon VPC).

planejamento de recursos empresariais (ERP)

Um sistema que automatiza e gerencia os principais processos de negócios (como contabilidade, MES e gerenciamento de projetos) para uma empresa.

criptografia envelopada

O processo de criptografar uma chave de criptografia com outra chave de criptografia. Para obter mais informações, consulte Criptografia de envelope na documentação AWS Key Management Service (AWS KMS).

ambiente

Uma instância de uma aplicação em execução. Estes são tipos comuns de ambientes na computação em nuvem:

- ambiente de desenvolvimento: uma instância de uma aplicação em execução que está disponível somente para a equipe principal responsável pela manutenção da aplicação. Ambientes de desenvolvimento são usados para testar mudanças antes de promovê-las para ambientes superiores. Esse tipo de ambiente às vezes é chamado de ambiente de teste.
- ambientes inferiores: todos os ambientes de desenvolvimento para uma aplicação, como aqueles usados para compilações e testes iniciais.
- ambiente de produção: uma instância de uma aplicação em execução que os usuários finais podem acessar. Em um CI/CD pipeline, o ambiente de produção é o último ambiente de implantação.
- ambientes superiores: todos os ambientes que podem ser acessados por usuários que não sejam a equipe principal de desenvolvimento. Isso pode incluir um ambiente de produção, ambientes de pré-produção e ambientes para testes de aceitação do usuário.

epic

Em metodologias ágeis, categorias funcionais que ajudam a organizar e priorizar seu trabalho. Os epics fornecem uma descrição de alto nível dos requisitos e das tarefas de implementação. Por exemplo, os épicos de segurança AWS da CAF incluem gerenciamento de identidade e acesso, controles de detetive, segurança de infraestrutura, proteção de dados e resposta a incidentes. Para obter mais informações sobre epics na estratégia de migração da AWS , consulte o guia de implementação do programa.

ERP

Veja planejamento de recursos empresariais.

análise exploratória de dados (EDA)

O processo de analisar um conjunto de dados para entender suas principais características. Você coleta ou agrega dados e, em seguida, realiza investigações iniciais para encontrar padrões, detectar anomalias e verificar suposições. O EDA é realizado por meio do cálculo de estatísticas resumidas e da criação de visualizações de dados.

F

tabela de fatos

A tabela central em um esquema em estrela. Ela armazena dados quantitativos sobre as operações comerciais. Normalmente, uma tabela de fatos contém dois tipos de colunas: as que contêm medidas e as que contêm uma chave externa para uma tabela de dimensões.

Antecipar-se à falha

Uma filosofia que usa testes frequentes e incrementais para reduzir o ciclo de vida do desenvolvimento. É uma parte essencial de uma abordagem ágil.

delimitação de isolamento contra falhas

No Nuvem AWS, um limite, como uma zona de disponibilidade, Região da AWS um plano de controle ou um plano de dados, que limita o efeito de uma falha e ajuda a melhorar a resiliência das cargas de trabalho. Para obter mais informações, consulte AWS Fault Isolation Boundaries.

ramificação de recursos

Veja ramificação.

recursos

Os dados de entrada usados para fazer uma previsão. Por exemplo, em um contexto de manufatura, os recursos podem ser imagens capturadas periodicamente na linha de fabricação.

importância do recurso

O quanto um recurso é importante para as previsões de um modelo. Isso geralmente é expresso como uma pontuação numérica que pode ser calculada por meio de várias técnicas, como Shapley Additive Explanations (SHAP) e gradientes integrados. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

transformação de recursos

O processo de otimizar dados para o processo de ML, incluindo enriquecer dados com fontes adicionais, escalar valores ou extrair vários conjuntos de informações de um único campo de dados. Isso permite que o modelo de ML se beneficie dos dados. Por exemplo, se a data “2021-05-27 00:15:37” for dividida em “2021”, “maio”, “quinta” e “15”, isso poderá ajudar o algoritmo de aprendizado a aprender padrões diferenciados associados a diferentes componentes de dados.

prompt few shot

Fornecer a um LLM um pequeno número de exemplos que demonstram a tarefa e o resultado desejado antes de solicitar que ele execute uma tarefa semelhante. Essa técnica é uma aplicação do aprendizado em contexto, em que os modelos aprendem com exemplos (shots) incorporados aos prompts. Prompts few-shot podem ser eficazes para tarefas que exigem formatação, raciocínio ou conhecimento de domínio específicos. Veja também prompts zero-shot.

FGAC

Veja controle de acesso refinado.

Controle de acesso refinado (FGAC)

O uso de várias condições para permitir ou negar uma solicitação de acesso.

migração flash-cut

Um método de migração de banco de dados que usa replicação contínua de dados via captura de dados de alteração para migrar os dados no menor tempo possível, em vez de usar uma abordagem em fases. O objetivo é reduzir ao mínimo o tempo de inatividade.

FM

Veja modelo de base.

modelo de base (FM)

Uma grande rede neural de aprendizado profundo que vem treinando em grandes conjuntos de dados generalizados e não rotulados. FMs são capazes de realizar uma ampla variedade de tarefas gerais, como entender a linguagem, gerar texto e imagens e conversar em linguagem natural. Para obter mais informações, consulte O que são modelos de base?.

G

IA generativa

Um subconjunto de modelos de IA que foram treinados em grandes quantidades de dados e que podem usar um simples prompt de texto para criar novos artefatos e conteúdo, como imagens, vídeos, texto e áudio. Para obter mais informações, consulte O que é IA generativa?.

bloqueio geográfico

Veja restrições geográficas.

restrições geográficas (bloqueio geográfico)

Na Amazon CloudFront, uma opção para impedir que usuários em países específicos acessem distribuições de conteúdo. É possível usar uma lista de permissões ou uma lista de bloqueios para especificar países aprovados e banidos. Para obter mais informações, consulte Restringir a distribuição geográfica do seu conteúdo na CloudFront documentação.

Fluxo de trabalho do GitFlow

Uma abordagem na qual ambientes inferiores e superiores usam ramificações diferentes em um repositório de código-fonte. O fluxo de trabalho do Gitflow é considerado legado, e o fluxo de trabalho trunk-based é a abordagem moderna e preferencial.

golden image

Um snapshot de um sistema ou software usado como modelo para implantar novas instâncias desse sistema ou software. Por exemplo, na manufatura, uma golden image pode ser usada para provisionar software em vários dispositivos e ajudar a melhorar a velocidade, a escalabilidade e a produtividade nas operações de fabricação de dispositivos.

estratégia greenfield

A ausência de infraestrutura existente em um novo ambiente. Ao adotar uma estratégia greenfield para uma arquitetura de sistema, é possível selecionar todas as novas tecnologias sem a restrição da compatibilidade com a infraestrutura existente, também conhecida como brownfield. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

barreira de proteção

Uma regra de alto nível que ajuda a governar recursos, políticas e conformidade em todas as unidades organizacionais ()OUs. Barreiras de proteção preventivas impõem políticas para garantir o alinhamento a padrões de conformidade. Elas são implementadas usando políticas de controle de serviço e limites de permissões do IAM. Barreiras de proteção detectivas detectam violações de políticas e problemas de conformidade e geram alertas para remediação. Eles são implementados usando AWS Config, AWS Security Hub CSPM, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector e verificações personalizadas AWS Lambda .

H

HA

Veja alta disponibilidade.

migração heterogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que usa um mecanismo de banco de dados diferente (por exemplo, Oracle para Amazon Aurora). A migração heterogênea geralmente faz parte de um esforço de redefinição da arquitetura, e converter

o esquema pode ser uma tarefa complexa. O AWS fornece o AWS SCT para ajudar nas conversões de esquemas.

alta disponibilidade (HA)

A capacidade de uma workload operar continuamente, sem intervenção, em caso de desafios ou desastres. Os sistemas AH são projetados para realizar o failover automático, oferecer consistentemente desempenho de alta qualidade e lidar com diferentes cargas e falhas com impacto mínimo no desempenho.

modernização de historiador

Uma abordagem usada para modernizar e atualizar os sistemas de tecnologia operacional (OT) para melhor atender às necessidades do setor de manufatura. Um historiador é um tipo de banco de dados usado para coletar e armazenar dados de várias fontes em uma fábrica.

dados de hold-out

Uma parte dos dados históricos rotulados que são retidos de um conjunto de dados usado para treinar um modelo de machine learning. Você pode usar dados de hold-out para avaliar a performance do modelo comparando as predições do modelo com os dados de retenção.

migração homogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que compartilha o mesmo mecanismo de banco de dados (por exemplo, Microsoft SQL Server para Amazon RDS para SQL Server). A migração homogênea geralmente faz parte de um esforço de redefinição da hospedagem ou da plataforma. É possível usar utilitários de banco de dados nativos para migrar o esquema.

dados quentes

Dados acessados com frequência, como dados em tempo real ou dados translacionais recentes. Esses dados normalmente exigem uma camada ou classe de armazenamento de alto desempenho para fornecer respostas rápidas às consultas.

hotfix

Uma correção urgente para um problema crítico em um ambiente de produção. Devido à sua urgência, um hotfix geralmente é feito fora do fluxo de trabalho normal de DevOps lançamento.

período de hipercuidados

Imediatamente após a substituição, o período em que uma equipe de migração gerencia e monitora as aplicações migradas na nuvem para resolver quaisquer problemas. Normalmente,

a duração desse período é de 1 a 4 dias. No final do período de hipercuidados, a equipe de migração normalmente transfere a responsabilidade pelas aplicações para a equipe de operações de nuvem.

eu

IaC

Veja infraestrutura como código.

Política baseada em identidade

Uma política anexada a um ou mais diretores do IAM que define suas permissões no Nuvem AWS ambiente.

aplicação ociosa

Uma aplicação que tem um uso médio de CPU e memória entre 5 e 20% em um período de 90 dias. Em um projeto de migração, é comum retirar essas aplicações ou retê-las on-premises.

IIoT

Veja Internet das Coisas Industrial.

infraestrutura imutável

Um modelo que implanta uma nova infraestrutura para workloads de produção em vez de atualizar, aplicar patches ou modificar a infraestrutura existente. Infraestruturas imutáveis são inerentemente mais consistentes, confiáveis e preditivas do que infraestruturas mutáveis. Para obter mais informações, consulte a prática recomendada Implantar usando infraestrutura imutável no AWS Well-Architected Framework.

VPC de entrada (admissão)

Em uma arquitetura de AWS várias contas, uma VPC que aceita, inspeciona e roteia conexões de rede de fora de um aplicativo. A Arquitetura de Referência de AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo e a Internet em geral.

migração incremental

Uma estratégia de substituição na qual você migra a aplicação em pequenas partes, em vez de realizar uma única substituição completa. Por exemplo, é possível mover inicialmente

apenas alguns microsserviços ou usuários para o novo sistema. Depois de verificar se tudo está funcionando corretamente, mova os microsserviços ou usuários adicionais de forma incremental até poder descomissionar seu sistema herdado. Essa estratégia reduz os riscos associados a migrações de grande porte.

Indústria 4.0

Um termo que foi introduzido por Klaus Schwab em 2016 para se referir à modernização dos processos de manufatura por meio de avanços em conectividade, dados em tempo real, automação, analytics e IA/ML.

infraestrutura

Todos os recursos e ativos contidos no ambiente de uma aplicação.

Infraestrutura como código (IaC)

O processo de provisionamento e gerenciamento da infraestrutura de uma aplicação por meio de um conjunto de arquivos de configuração. A IaC foi projetada para ajudar você a centralizar o gerenciamento da infraestrutura, padronizar recursos e escalar rapidamente para que novos ambientes sejam reproduzíveis, confiáveis e consistentes.

Internet industrial das coisas (IIoT)

O uso de sensores e dispositivos conectados à Internet nos setores industriais, como manufatura, energia, automotivo, saúde, ciências biológicas e agricultura. Para obter mais informações, consulte Criando uma estratégia de transformação digital industrial da Internet das Coisas (IIoT).

VPC de inspeção

Em uma arquitetura de AWS várias contas, uma VPC centralizada que gerencia as inspeções do tráfego de rede entre VPCs (na mesma ou em diferentes Regiões da AWS) a Internet e as redes locais. A Arquitetura de Referência de AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo e a Internet em geral.

Internet das coisas (IoT)

A rede de objetos físicos conectados com sensores ou processadores incorporados que se comunicam com outros dispositivos e sistemas pela Internet ou por uma rede de comunicação local. Para obter mais informações, consulte O que é IoT?

interpretabilidade

Uma característica de um modelo de machine learning que descreve o grau em que um ser humano pode entender como as previsões do modelo dependem de suas entradas. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

IoT

Veja Internet das Coisas.

Biblioteca de informações de TI (ITIL)

Um conjunto de práticas recomendadas para fornecer serviços de TI e alinhar esses serviços a requisitos de negócios. A ITIL fornece a base para o ITSM.

Gerenciamento de serviços de TI (ITSM)

Atividades associadas a design, implementação, gerenciamento e suporte de serviços de TI para uma organização. Para obter informações sobre a integração de operações em nuvem com ferramentas de ITSM, consulte o guia de integração de operações.

ITIL

Veja biblioteca de informações de TI.

ITSM

Veja gerenciamento de serviços de TI.

L

controle de acesso baseado em etiqueta (LBAC)

Uma implementação do controle de acesso obrigatório (MAC) em que os usuários e os dados em si recebem explicitamente um valor de etiqueta de segurança. A interseção entre a etiqueta de segurança do usuário e a etiqueta de segurança dos dados determina quais linhas e colunas podem ser vistas pelo usuário.

zona de pouso

Uma landing zone é um AWS ambiente bem arquitetado, com várias contas, escalável e seguro. Um ponto a partir do qual suas organizações podem iniciar e implantar rapidamente workloads e aplicações com confiança em seu ambiente de segurança e infraestrutura. Para obter mais

informações sobre zonas de pouso, consulte Configurar um ambiente da AWS com várias contas seguro e escalável.

grande modelo de linguagem (LLM)

Um modelo de IA de aprendizado profundo pré-treinado em uma grande quantidade de dados. Um LLM pode realizar várias tarefas, como responder a perguntas, resumir documentos, traduzir texto para outros idiomas e completar frases. Para obter mais informações, consulte O que são LLMs.

migração de grande porte

Uma migração de 300 servidores ou mais.

LBAC

Veja controle de acesso baseado em rótulo.

privilégio mínimo

A prática recomendada de segurança de conceder as permissões mínimas necessárias para executar uma tarefa. Para obter mais informações, consulte Aplicar permissões de privilégios mínimos na documentação do IAM.

mover sem alterações (lift-and-shift)

Veja 7 Rs.

sistema little-endian

Um sistema que armazena o byte menos significativo antes. Veja também endianness.

LLM

Veja grande modelo de linguagem.

ambientes inferiores

Veja ambiente.

M

machine learning (ML)

Um tipo de inteligência artificial que usa algoritmos e técnicas para reconhecimento e aprendizado de padrões. O ML analisa e aprende com dados gravados, por exemplo, dados da

Internet das Coisas (IoT), para gerar um modelo estatístico baseado em padrões. Para obter mais informações, consulte Machine learning.

ramificação principal

Veja ramificação.

Malware

Software projetado para comprometer a segurança ou a privacidade do computador. O malware pode interromper os sistemas do computador, vazar informações sensíveis ou obter acesso não autorizado. Exemplos de malware incluem vírus, worms, ransomware, cavalos de Troia, spyware e keyloggers.

Serviços gerenciados

Serviços da AWS para o qual AWS opera a camada de infraestrutura, o sistema operacional e as plataformas, e você acessa os endpoints para armazenar e recuperar dados. O Amazon Simple Storage Service (Amazon S3) e o Amazon DynamoDB são exemplos de serviços gerenciados. Eles também são conhecidos como serviços abstraídos.

sistema de execução de manufatura (MES)

Um sistema de software para rastrear, monitorar, documentar e controlar processos de produção que convertem matérias-primas em produtos acabados no chão de fábrica.

MAP

Veja Programa de Aceleração da Migração.

mecanismo

Um processo completo em que você cria uma ferramenta, impulsiona a adoção da ferramenta e, em seguida, inspeciona os resultados para fazer ajustes. Um mecanismo é um ciclo que se reforça e se aprimora à medida que opera. Para obter mais informações, consulte Construindo mecanismos no AWS Well-Architected Framework.

conta de membro

Todos, Contas da AWS exceto a conta de gerenciamento, que fazem parte de uma organização em AWS Organizations. Uma conta só pode ser membro de uma organização de cada vez.

MES

Veja sistema de execução de manufatura.

Transporte de Telemetria de Enfileiramento de Mensagens (MQTT)

Um protocolo de comunicação leve machine-to-machine (M2M), baseado no padrão de publicação/assinatura, para dispositivos de IoT com recursos limitados.

microsserviço

Um serviço pequeno e independente que se comunica de forma bem definida APIs e normalmente é de propriedade de equipes pequenas e independentes. Por exemplo, um sistema de seguradora pode incluir microsserviços que mapeiam as capacidades comerciais, como vendas ou marketing, ou subdomínios, como compras, reclamações ou análises. Os benefícios dos microsserviços incluem agilidade, escalabilidade flexível, fácil implantação, código reutilizável e resiliência. Para obter mais informações, consulte Integração de microsserviços usando serviços sem AWS servidor.

arquitetura de microsserviços

Uma abordagem à criação de aplicações com componentes independentes que executam cada processo de aplicação como um microsserviço. Esses microsserviços se comunicam por meio de uma interface bem definida usando leveza. APIs Cada microsserviço nessa arquitetura pode ser atualizado, implantado e escalado para atender à demanda por funções específicas de uma aplicação. Para obter mais informações, consulte Implementação de microsserviços em. AWS

Programa de Aceleração da Migração (MAP)

Um AWS programa que fornece suporte de consultoria, treinamento e serviços para ajudar as organizações a criar uma base operacional sólida para migrar para a nuvem e ajudar a compensar o custo inicial das migrações. O MAP inclui uma metodologia de migração para executar migrações legadas de forma metódica e um conjunto de ferramentas para automatizar e acelerar cenários comuns de migração.

migração em escala

O processo de mover a maior parte do portfólio de aplicações para a nuvem em ondas, com mais aplicações sendo movidas em um ritmo mais rápido a cada onda. Essa fase usa as práticas recomendadas e lições aprendidas nas fases anteriores para implementar uma fábrica de migração de equipes, ferramentas e processos para agilizar a migração de workloads por meio de automação e entrega ágeis. Esta é a terceira fase da estratégia de migração para a AWS.

fábrica de migração

Equipes multifuncionais que simplificam a migração de workloads por meio de abordagens automatizadas e ágeis. As equipes da fábrica de migração geralmente incluem operações,

analistas e proprietários de negócios, engenheiros de migração, desenvolvedores e DevOps profissionais que trabalham em sprints. Entre 20 e 50% de um portfólio de aplicações corporativas consiste em padrões repetidos que podem ser otimizados por meio de uma abordagem de fábrica. Para obter mais informações, consulte discussão sobre fábricas de migração e o guia do Cloud Migration Factory neste conjunto de conteúdo.

metadados de migração

As informações sobre a aplicação e o servidor necessárias para concluir a migração. Cada padrão de migração exige um conjunto de metadados de migração diferente. Exemplos de metadados de migração incluem a sub-rede, o grupo de segurança e AWS a conta de destino.

padrão de migração

Uma tarefa de migração repetível que detalha a estratégia de migração, o destino da migração e a aplicação ou o serviço de migração usado. Exemplo: rehospede a migração para o Amazon EC2 AWS com o Application Migration Service.

Avaliação de Portfólio para Migração (MPA)

Uma ferramenta on-line que fornece informações para validar o caso de negócios para migrar para a Nuvem AWS. O MPA fornece avaliação detalhada do portfólio (dimensionamento correto do servidor, preços, comparações de TCO, análise de custos de migração), bem como planejamento de migração (análise e coleta de dados de aplicações, agrupamento de aplicações, priorização de migração e planejamento de ondas). A ferramenta MPA (requer login) está disponível gratuitamente para todos os AWS consultores e consultores parceiros da APN.

Avaliação de Preparação para Migração (MRA)

O processo de obter insights sobre o status de prontidão de uma organização para a nuvem, identificar pontos fortes e fracos e criar um plano de ação para fechar as lacunas identificadas, usando o CAF. AWS Para mais informações, consulte o guia de preparação para migração. A MRA é a primeira fase da estratégia de migração para a AWS.

estratégia de migração

A abordagem usada para migrar uma workload para a Nuvem AWS. Para obter mais informações, veja a entrada 7 Rs neste glossário e consulte Mobilize sua organização para acelerar migrações em grande escala.

ML

Veja machine learning.

modernização

Transformar uma aplicação desatualizada (herdada ou monolítica) e sua infraestrutura em um sistema ágil, elástico e altamente disponível na nuvem para reduzir custos, ganhar eficiência e aproveitar as inovações. Para obter mais informações, consulte Strategy for modernizing applications in the Nuvem AWS.

avaliação de preparação para modernização

Uma avaliação que ajuda a determinar a preparação para modernização das aplicações de uma organização. Ela identifica benefícios, riscos e dependências e determina o quão bem a organização pode acomodar o estado futuro dessas aplicações. O resultado da avaliação é um esquema da arquitetura de destino, um roteiro que detalha as fases de desenvolvimento e os marcos do processo de modernização e um plano de ação para abordar as lacunas identificadas. Para obter mais informações, consulte Evaluating modernization readiness for applications in the Nuvem AWS.

aplicações monolíticas (monólitos)

Aplicações que são executadas como um único serviço com processos fortemente acoplados. As aplicações monolíticas apresentam várias desvantagens. Se um recurso da aplicação apresentar um aumento na demanda, toda a arquitetura deverá ser escalada. Adicionar ou melhorar os recursos de uma aplicação monolítica também se torna mais complexo quando a base de código cresce. Para resolver esses problemas, é possível criar uma arquitetura de microsserviços. Para obter mais informações, consulte Decompor monólitos em microsserviços.

MPA

Veja Avaliação do Portfólio para Migração.

MQTT

Veja Transporte de Telemetria de Enfileiramento de Mensagens.

classificação multiclasse

Um processo que ajuda a gerar previsões para várias classes (prevendo um ou mais de dois resultados). Por exemplo, um modelo de ML pode perguntar “Este produto é um livro, um carro ou um telefone?” ou “Qual categoria de produtos é mais interessante para este cliente?”

infraestrutura mutável

Um modelo que atualiza e modifica a infraestrutura existente para workloads de produção. Para melhorar a consistência, confiabilidade e previsibilidade, o AWS Well-Architected Framework recomenda o uso de infraestrutura imutável como uma prática recomendada.

O

OAC

Veja controle de acesso de origem.

OAI

Veja identidade de acesso de origem.

OCM

Veja gerenciamento de alterações organizacionais.

migração offline

Um método de migração no qual a workload de origem é desativada durante o processo de migração. Esse método envolve tempo de inatividade prolongado e geralmente é usado para workloads pequenas e não críticas.

OI

Veja integração de operações.

Ola

Veja acordo de nível operacional.

migração online

Um método de migração no qual a workload de origem é copiada para o sistema de destino sem ser colocada offline. As aplicações conectadas à workload podem continuar funcionando durante a migração. Esse método envolve um tempo de inatividade nulo ou mínimo e normalmente é usado para workloads essenciais para a produção.

OPC-UA

Veja Open Process Communications - Unified Architecture.

Open Process Communications - Unified Architecture (OPC-UA)

Um protocolo de comunicação machine-to-machine (M2M) para automação industrial. O OPC-UA fornece um padrão de interoperabilidade com esquemas de criptografia, autenticação e autorização de dados.

acordo de nível operacional (OLA)

Um acordo que esclarece o que os grupos funcionais de TI prometem oferecer uns aos outros para apoiar um acordo de serviço (SLA).

análise de prontidão operacional (ORR)

Uma lista de verificação de perguntas e práticas recomendadas associadas que ajudam você a entender, avaliar, prevenir ou reduzir o escopo de incidentes e possíveis falhas. Para obter mais informações, consulte Operational Readiness Reviews (ORR) no AWS Well-Architected Framework.

tecnologia operacional (TO)

Sistemas de hardware e software que trabalham com o ambiente físico para controlar operações, equipamentos e infraestrutura industriais. Na manufatura, a integração dos sistemas de tecnologia da informação (TI) e tecnologia operacional (TO) é o foco principal das transformações da Indústria 4.0.

integração de operações (OI)

O processo de modernização das operações na nuvem, que envolve planejamento de preparação, automação e integração. Para obter mais informações, consulte o guia de integração de operações.

trilha organizacional

Uma trilha criada por ela AWS CloudTrail registra todos os eventos de todos Contas da AWS em uma organização em AWS Organizations. Essa trilha é criada em cada Conta da AWS que faz parte da organização e monitora a atividade em cada conta. Para obter mais informações, consulte Criação de uma trilha para uma organização na CloudTrail documentação.

gerenciamento de alterações organizacionais (OCM)

Uma estrutura para gerenciar grandes transformações de negócios disruptivas de uma perspectiva de pessoas, cultura e liderança. O OCM ajuda as organizações a se prepararem e fazerem a transição para novos sistemas e estratégias, acelerando a adoção de alterações, abordando questões de transição e promovendo mudanças culturais e organizacionais. Na estratégia de AWS migração, essa estrutura é chamada de aceleração de pessoas, devido à velocidade de mudança exigida nos projetos de adoção da nuvem. Para obter mais informações, consulte o guia do OCM.

controle de acesso de origem (OAC)

Em CloudFront, uma opção aprimorada para restringir o acesso para proteger seu conteúdo do Amazon Simple Storage Service (Amazon S3). O OAC oferece suporte a todos os buckets S3 Regiões da AWS, criptografia do lado do servidor com AWS KMS (SSE-KMS) e solicitações dinâmicas ao bucket S3. `PUT` `DELETE`

Identidade do acesso de origem (OAI)

Em CloudFront, uma opção para restringir o acesso para proteger seu conteúdo do Amazon S3. Quando você usa o OAI, CloudFront cria um principal com o qual o Amazon S3 pode se autenticar. Os diretores autenticados podem acessar o conteúdo em um bucket do S3 somente por meio de uma distribuição específica. CloudFront Veja também OAC, que fornece um controle de acesso mais granular e aprimorado.

ORR

Veja análise de prontidão operacional.

OT

Veja tecnologia operacional.

VPC de saída (egresso)

Em uma arquitetura de AWS várias contas, uma VPC que gerencia conexões de rede que são iniciadas de dentro de um aplicativo. A Arquitetura de Referência de AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo e a Internet em geral.

P

limite de permissões

Uma política de gerenciamento do IAM anexada a entidades principais do IAM para definir as permissões máximas que o usuário ou perfil podem ter. Para obter mais informações, consulte Limites de permissões na documentação do IAM.

Informações de identificação pessoal (PII)

Informações que, quando visualizadas diretamente ou combinadas com outros dados relacionados, podem ser usadas para inferir razoavelmente a identidade de um indivíduo. Exemplos de PII incluem nomes, endereços e informações de contato.

PII

Veja informações de identificação pessoal.

manual

Um conjunto de etapas predefinidas que capturam o trabalho associado às migrações, como a entrega das principais funções operacionais na nuvem. Um manual pode assumir a forma de scripts, runbooks automatizados ou um resumo dos processos ou etapas necessários para operar seu ambiente modernizado.

PLC

Veja controlador lógico programável.

PLM

Veja gerenciamento do ciclo de vida do produto.

política

Um objeto que pode definir permissões (veja política baseada em identidade), especificar condições de acesso (veja políta baseada em recurso) ou definir as permissões máximas para todas as contas em uma organização no AWS Organizations (veja política de controle de serviços).

persistência poliglota

Escolher de forma independente a tecnologia de armazenamento de dados de um microsserviço com base em padrões de acesso a dados e outros requisitos. Se seus microsserviços tiverem a mesma tecnologia de armazenamento de dados, eles poderão enfrentar desafios de implementação ou apresentar baixa performance. Os microsserviços serão implementados com mais facilidade e alcançarão performance e escalabilidade melhores se usarem o armazenamento de dados mais bem adaptado às suas necessidades.

avaliação do portfólio

Um processo de descobrir, analisar e priorizar o portfólio de aplicações para planejar a migração. Para obter mais informações, consulte Avaliar a preparação para a migração.

predicado

Uma condição de consulta que retorna `true` ou `false`, normalmente localizada em uma cláusula `WHERE`.

pushdown de predicados

Uma técnica de otimização de consultas de banco de dados que filtra os dados na consulta antes da transferência. Isso reduz a quantidade de dados que devem ser recuperados e processados do banco de dados relacional e melhora a performance das consultas.

controle preventivo

Um controle de segurança projetado para evitar que um evento ocorra. Esses controles são a primeira linha de defesa para ajudar a evitar acesso não autorizado ou alterações indesejadas em sua rede. Para obter mais informações, consulte Controles preventivos em Como implementar controles de segurança na AWS.

principal (entidade principal)

Uma entidade AWS que pode realizar ações e acessar recursos. Essa entidade geralmente é um usuário raiz para um Conta da AWS, uma função do IAM ou um usuário. Para obter mais informações, consulte Entidade principal em Termos e conceitos de perfis na documentação do IAM.

Privacidade por design

Uma abordagem em engenharia de sistemas que leva em consideração a privacidade em todo o processo de desenvolvimento.

zonas hospedadas privadas

Um contêiner que contém informações sobre como você deseja que o Amazon Route 53 responda às consultas de DNS para um domínio e seus subdomínios em um ou mais. VPCs Para obter mais informações, consulte Como trabalhar com zonas hospedadas privadas na documentação do Route 53.

controle proativo

Um controle de segurança desenvolvido para evitar a implantação de recursos não conformes. Esses controles verificam os recursos antes de serem provisionados. Se o recurso não estiver em conformidade com o controle, ele não será provisionado. Para obter mais informações, consulte o guia de referência de controles na AWS Control Tower documentação e consulte Controles proativos em Implementação de controles de segurança em AWS.

gerenciamento do ciclo de vida do produto (PLM)

O gerenciamento de dados e processos de um produto em todo o seu ciclo de vida, desde a concepção, o desenvolvimento e o lançamento, passando pelo crescimento e maturidade, até o declínio e a remoção.

ambiente de produção

Veja ambiente.

controlador lógico programável (PLC)

Na manufatura, um computador altamente confiável e adaptável que monitora as máquinas e automatiza os processos de fabricação.

encadeamento de prompts

Uso da saída de um prompt do LLM como entrada para o próximo prompt para gerar respostas melhores. Essa técnica é usada para dividir uma tarefa complexa em subtarefas, ou para refinar ou expandir iterativamente uma resposta preliminar. Isso ajuda a melhorar a precisão e a relevância das respostas de um modelo e permite resultados mais granulares e personalizados.

pseudonimização

O processo de substituir identificadores pessoais em um conjunto de dados por valores de espaço reservado. A pseudonimização pode ajudar a proteger a privacidade pessoal. Os dados pseudonimizados ainda são considerados dados pessoais.

publish/subscribe (pub/sub)

Um padrão que permite comunicações assíncronas entre microsserviços para melhorar a escalabilidade e a capacidade de resposta. Por exemplo, em um MES baseado em microsserviços, um microsserviço pode publicar mensagens de eventos em um canal em que outros microsserviços possam assinar. O sistema pode adicionar novos microsserviços sem alterar o serviço de publicação.

Q

plano de consulta

Uma série de etapas, como instruções, usadas para acessar os dados em um sistema de banco de dados relacional SQL.

regressão de planos de consultas

Quando um otimizador de serviço de banco de dados escolhe um plano menos adequado do que escolhia antes de uma determinada alteração no ambiente de banco de dados ocorrer. Isso pode ser causado por alterações em estatísticas, restrições, configurações do ambiente, associações de parâmetros de consulta e atualizações do mecanismo de banco de dados.

R

Matriz RACI

Veja responsável, aprovador, consultado, informado (RACI).

RAG

Veja geração aumentada via recuperação.

ransomware

Um software mal-intencionado desenvolvido para bloquear o acesso a um sistema ou dados de computador até que um pagamento seja feito.

Matriz RASCI

Veja responsável, aprovador, consultado, informado (RACI).

RCAC

Veja controle de acesso por linha e coluna.

réplica de leitura

Uma cópia de um banco de dados usada somente para leitura. É possível encaminhar consultas para a réplica de leitura e reduzir a carga no banco de dados principal.

Redefinir arquitetura

Veja 7 Rs.

objetivo de ponto de recuperação (RPO).

O máximo período de tempo 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.

objetivo de tempo de recuperação (RTO)

O máximo atraso aceitável entre a interrupção e a restauração do serviço.

refatorar

Veja 7 Rs.

Região

Uma coleção de AWS recursos em uma área geográfica. Cada um Região da AWS é isolado e independente dos outros para fornecer tolerância a falhas, estabilidade e resiliência. Para obter informações, consulte Specify which Regiões da AWS your account can use.

regressão

Uma técnica de ML que prevê um valor numérico. Por exemplo, para resolver o problema de “Por qual preço esta casa será vendida?” um modelo de ML pode usar um modelo de regressão linear para prever o preço de venda de uma casa com base em fatos conhecidos sobre a casa (por exemplo, a metragem quadrada).

redefinir a hospedagem

Veja 7 Rs.

versão

Em um processo de implantação, o ato de promover mudanças em um ambiente de produção.

realocar

Veja 7 Rs.

redefinir a plataforma

Veja 7 Rs.

recomprar

Veja 7 Rs.

resiliência

A capacidade de uma aplicação de resistir ou se recuperar de interrupções. Alta disponibilidade e recuperação de desastres são considerações comuns ao planejar a resiliência na Nuvem AWS. Para obter mais informações, consulte Nuvem AWS Resilience.

política baseada em recurso

Uma política associada a um recurso, como um bucket do Amazon S3, um endpoint ou uma chave de criptografia. Esse tipo de política especifica quais entidades principais têm acesso permitido, ações válidas e quaisquer outras condições que devem ser atendidas.

matriz responsável, accountable, consultada, informada (RACI)

Uma matriz que define as funções e responsabilidades de todas as partes envolvidas nas atividades de migração e nas operações de nuvem. O nome da matriz é derivado dos tipos de responsabilidade definidos na matriz: responsável (R), responsabilizável (A), consultado (C) e informado (I). O tipo de suporte (S) é opcional. Se você incluir suporte, a matriz será chamada de matriz RASCI e, se excluir, será chamada de matriz RACI.

controle responsivo

Um controle de segurança desenvolvido para conduzir a remediação de eventos adversos ou desvios em relação à linha de base de segurança. Para obter mais informações, consulte Controles responsivos em Como implementar controles de segurança na AWS.

reter

Veja 7 Rs.

Retirada

Veja 7 Rs.

Geração Aumentada de Recuperação (RAG)

Uma tecnologia de IA generativa em que um LLM faz referência a uma fonte de dados autorizada que está fora de suas fontes de dados de treinamento antes de gerar uma resposta. Por exemplo, um modelo RAG pode realizar uma pesquisa semântica na base de conhecimento ou nos dados personalizados de uma organização. Para obter mais informações, consulte O que é RAG (geração aumentada via recuperação)?.

alternância

O processo de atualizar periodicamente um segredo para dificultar o acesso de um invasor às credenciais.

controle de acesso por linha e coluna (RCAC)

O uso de expressões SQL básicas e flexíveis que tenham regras de acesso definidas. O RCAC consiste em permissões de linha e máscaras de coluna.

RPO

Veja objetivo de ponto de recuperação.

RTO

Veja objetivo de tempo de recuperação.

runbook

Um conjunto de procedimentos manuais ou automatizados necessários para realizar uma tarefa específica. Eles são normalmente criados para agilizar operações ou procedimentos repetitivos com altas taxas de erro.

S

SAML 2.0

Um padrão aberto que muitos provedores de identidade (IdPs) usam. Esse recurso permite o login único federado (SSO), para que os usuários possam fazer login Console de gerenciamento da AWS ou chamar as operações da AWS API sem que você precise criar um usuário no IAM para todos em sua organização. Para obter mais informações sobre a federação baseada em SAML 2.0, consulte Sobre a federação baseada em SAML 2.0 na documentação do IAM.

SCADA

Veja controle de supervisão e aquisição de dados.

SCP

Veja política de controle de serviço.

secret

Em AWS Secrets Manager, informações confidenciais ou restritas, como uma senha ou credenciais de usuário, que você armazena de forma criptografada. Consiste no valor secreto e em seus metadados. O valor secreto pode ser binário, uma única string ou várias strings. Para obter mais informações, consulte What's in a Secrets Manager secret? na documentação do Secrets Manager.

segurança desde a concepção

Uma abordagem em engenharia de sistemas que leva em consideração a segurança em todo o processo de desenvolvimento.

controle de segurança

Uma barreira de proteção técnica ou administrativa que impede, detecta ou reduz a capacidade de uma ameaça explorar uma vulnerabilidade de segurança. Existem quatro tipos primários de controles de segurança: preventivos, detectivos, responsivos e proativos.

hardening da segurança

O processo de reduzir a superfície de ataque para torná-la mais resistente a ataques. Isso pode incluir ações como remover recursos que não são mais necessários, implementar a prática recomendada de segurança de conceder privilégios mínimos ou desativar recursos desnecessários em arquivos de configuração.

sistema de gerenciamento de eventos e informações de segurança (SIEM)

Ferramentas e serviços que combinam sistemas de gerenciamento de informações de segurança (SIM) e gerenciamento de eventos de segurança (SEM). Um sistema SIEM coleta, monitora e analisa dados de servidores, redes, dispositivos e outras fontes para detectar ameaças e violações de segurança e gerar alertas.

automação de resposta de segurança

Uma ação predefinida e programada projetada para responder ou remediar automaticamente um evento de segurança. Essas automações servem como controles de segurança responsivos ou detectivos que ajudam você a implementar as melhores práticas AWS de segurança. Exemplos de ações de resposta automatizada incluem a modificação de um grupo de segurança da VPC, a aplicação de patches em uma instância do Amazon EC2 ou a alternância de credenciais.

Criptografia do lado do servidor

Criptografia dos dados em seu destino, por AWS service (Serviço da AWS) quem os recebe.

política de controle de serviços (SCP)

Uma política que fornece controle centralizado sobre as permissões de todas as contas em uma organização em AWS Organizations. SCPs defina barreiras ou estabeleça limites nas ações que um administrador pode delegar a usuários ou funções. Você pode usar SCPs como listas de permissão ou listas de negação para especificar quais serviços ou ações são permitidos ou proibidos. Para obter mais informações, consulte Políticas de controle de serviço na AWS Organizations documentação.

service endpoint (endpoint de serviço)

O URL do ponto de entrada para um AWS service (Serviço da AWS). Você pode usar o endpoint para se conectar programaticamente ao serviço de destino. Para obter mais informações, consulte Endpoints do AWS service (Serviço da AWS) na Referência geral da AWS.

acordo de serviço (SLA)

Um acordo que esclarece o que uma equipe de TI promete fornecer aos clientes, como tempo de atividade e performance do serviço.

indicador de nível de serviço (SLI)

Uma avaliação de um aspecto de performance de um serviço, como taxa de erro, disponibilidade ou throughput.

objetivo de nível de serviço (SLO)

Uma métrica alvo que representa a integridade de um serviço, conforme avaliado por um indicador de nível de serviço.

modelo de responsabilidade compartilhada

Um modelo que descreve a responsabilidade com a qual você compartilha AWS pela segurança e conformidade na nuvem. AWS é responsável pela segurança da nuvem, enquanto você é responsável pela segurança na nuvem. Para obter mais informações, consulte o Modelo de responsabilidade compartilhada.

SIEM

Veja sistema de gerenciamento de eventos e informações de segurança.

ponto único de falha (SPOF)

Uma falha em um único componente crítico de uma aplicação que pode interromper o sistema.

SLA

Veja cacordo de serviço.

SLI

Veja indicador de nível de serviço.

SLO

Veja objetivo de nível de serviço.

split-and-seed modelo

Um padrão para escalar e acelerar projetos de modernização. À medida que novos recursos e lançamentos de produtos são definidos, a equipe principal se divide para criar novas equipes de produtos. Isso ajuda a escalar os recursos e os serviços da sua organização, melhora a produtividade do desenvolvedor e possibilita inovações rápidas. Para obter mais informações, consulte Phased approach to modernizing applications in the Nuvem AWS.

SPOF

Veja ponto único de falha.

esquema em estrela

Uma estrutura organizacional de banco de dados que usa uma grande tabela de fatos para armazenar dados transacionais ou medidos e usa uma ou mais tabelas dimensionais menores para armazenar atributos de dados. Essa estrutura foi projetada para ser usada em um data warehouse ou para fins de inteligência comercial.

padrão strangler fig

Uma abordagem à modernização de sistemas monolíticos que consiste em reescrever e substituir incrementalmente a funcionalidade do sistema até que o sistema herdado possa ser desativado. Esse padrão usa a analogia de uma videira que cresce e se torna uma árvore estabelecida e, eventualmente, supera e substitui sua hospedeira. O padrão foi apresentado por Martin Fowler como forma de gerenciar riscos ao reescrever sistemas monolíticos. Para ver um exemplo de como aplicar esse padrão, consulte Modernizar incrementalmente os serviços Web herdados do Microsoft ASP.NET (ASMX) usando contêineres e o Amazon API Gateway.

sub-rede

Um intervalo de endereços IP na VPC. Cada sub-rede fica alocada em uma única zona de disponibilidade.

controle supervisório e aquisição de dados (SCADA)

Na manufatura, um sistema que usa hardware e software para monitorar ativos físicos e operações de produção.

symmetric encryption (criptografia simétrica)

Um algoritmo de criptografia que usa a mesma chave para criptografar e descriptografar dados.

testes sintéticos

Testar um sistema de forma que simule as interações do usuário para detectar possíveis problemas ou monitorar a performance. Você pode usar o Amazon CloudWatch Synthetics para criar esses testes.

prompt do sistema

Uma técnica para fornecer contexto, instruções ou orientações a um LLM a fim de direcionar seu comportamento. Os prompts do sistema ajudam a definir o contexto e a estabelecer regras para interações com os usuários.

T

tags

Pares de valores-chave que atuam como metadados para organizar seus recursos. AWS As tags podem ajudar você a gerenciar, identificar, organizar, pesquisar e filtrar recursos da . Para obter mais informações, consulte Marcar seus recursos do AWS.

variável-alvo

O valor que você está tentando prever no ML supervisionado. Ela também é conhecida como variável de resultado. Por exemplo, em uma configuração de fabricação, a variável-alvo pode ser um defeito do produto.

lista de tarefas

Uma ferramenta usada para monitorar o progresso por meio de um runbook. Uma lista de tarefas contém uma visão geral do runbook e uma lista de tarefas gerais a serem concluídas. Para cada tarefa geral, ela inclui o tempo estimado necessário, o proprietário e o progresso.

ambiente de teste

Veja ambiente.

treinamento

O processo de fornecer dados para que seu modelo de ML aprenda. Os dados de treinamento devem conter a resposta correta. O algoritmo de aprendizado descobre padrões nos dados de treinamento que mapeiam os atributos dos dados de entrada no destino (a resposta que você deseja prever). Ele gera um modelo de ML que captura esses padrões. Você pode usar o modelo de ML para obter previsões de novos dados cujo destino você não conhece.

gateway de trânsito

Um hub de trânsito de rede que você pode usar para interconectar sua rede com VPCs a rede local. Para obter mais informações, consulte O que é um gateway de trânsito na AWS Transit Gateway documentação.

fluxo de trabalho baseado em troncos

Uma abordagem na qual os desenvolvedores criam e testam recursos localmente em uma ramificação de recursos e, em seguida, mesclam essas alterações na ramificação principal. A ramificação principal é então criada para os ambientes de desenvolvimento, pré-produção e produção, sequencialmente.

Acesso confiável

Conceder permissões a um serviço que você especifica para realizar tarefas em sua organização AWS Organizations e em suas contas em seu nome. O serviço confiável cria um perfil vinculado ao serviço em cada conta, quando esse perfil é necessário, para realizar tarefas de gerenciamento para você. Para obter mais informações, consulte Usando AWS Organizations com outros AWS serviços na AWS Organizations documentação.

tuning (ajustar)

Alterar aspectos do processo de treinamento para melhorar a precisão do modelo de ML. Por exemplo, você pode treinar o modelo de ML gerando um conjunto de rótulos, adicionando rótulos e repetindo essas etapas várias vezes em configurações diferentes para otimizar o modelo.

equipe de duas pizzas

Uma pequena DevOps equipe que você pode alimentar com duas pizzas. Uma equipe de duas pizzas garante a melhor oportunidade possível de colaboração no desenvolvimento de software.

U

incerteza

Um conceito que se refere a informações imprecisas, incompletas ou desconhecidas que podem minar a confiabilidade dos modelos preditivos de ML. Há dois tipos de incertezas: a incerteza epistêmica é causada por dados limitados e incompletos, enquanto a incerteza aleatória é causada pelo ruído e pela aleatoriedade inerentes aos dados.

tarefas indiferenciadas

Também conhecido como trabalho pesado, trabalho necessário para criar e operar um aplicativo, mas que não fornece valor direto ao usuário final nem oferece vantagem competitiva. Exemplos de tarefas indiferenciadas incluem aquisição, manutenção e planejamento de capacidade.

ambientes superiores

Veja ambiente.

V

aspiração

Uma operação de manutenção de banco de dados que envolve limpeza após atualizações incrementais para recuperar armazenamento e melhorar a performance.

controle de versões

Processos e ferramentas que rastreiam mudanças, como alterações no código-fonte em um repositório.

emparelhamento da VPC

Uma conexão entre duas VPCs que permite rotear o tráfego usando endereços IP privados. Para ter mais informações, consulte O que é emparelhamento de VPC? na documentação da Amazon VPC.

Vulnerabilidade

Uma falha de software ou hardware que compromete a segurança do sistema.

W

cache quente

Um cache de buffer que contém dados atuais e relevantes que são acessados com frequência. A instância do banco de dados pode ler do cache do buffer, o que é mais rápido do que ler da memória principal ou do disco.

dados mornos

Dados acessados raramente. Ao consultar esse tipo de dados, consultas moderadamente lentas geralmente são aceitáveis.

função de janela

Uma função SQL que executa um cálculo em um grupo de linhas que se relacionam de alguma forma com o registro atual. As funções de janela são úteis para processar tarefas, como calcular uma média móvel ou acessar o valor das linhas com base na posição relativa da linha atual.

workload

Uma coleção de códigos e recursos que geram valor empresarial, como uma aplicação voltada para o cliente ou um processo de backend.

workstreams

Grupos funcionais em um projeto de migração que são responsáveis por um conjunto específico de tarefas. Cada workstream é independente, mas oferece suporte aos outros workstreams do projeto. Por exemplo, o workstream de portfólio é responsável por priorizar aplicações, planejar ondas e coletar metadados de migração. O workstream de portfólio entrega esses ativos ao workstream de migração, que então migra os servidores e as aplicações.

WORM

Veja gravação única e várias leituras.

WQF

Veja AWS Workload Qualification Framework.

gravação única e várias leituras (WORM)

Um modelo de armazenamento que grava dados uma única vez e evita que os dados sejam excluídos ou modificados. Os usuários autorizados podem ler os dados quantas vezes forem necessárias, mas não podem alterá-los. Essa infraestrutura de armazenamento de dados é considerada imutável.

Z

exploração de dia zero

Um ataque, normalmente malware, que tira proveito de uma vulnerabilidade zero-day.

vulnerabilidade de dia zero

Uma falha ou vulnerabilidade não mitigada em um sistema de produção. Os agentes de ameaças podem usar esse tipo de vulnerabilidade para atacar o sistema. Os desenvolvedores frequentemente ficam cientes da vulnerabilidade como resultado do ataque.

prompt zero shot

Fornecer a um LLM instruções para realizar uma tarefa, mas sem exemplos (shots) que possam ajudar a orientá-lo. O LLM deve usar seu conhecimento pré-treinado para lidar com a tarefa. A

eficácia dos prompts zero-shot depende da complexidade da tarefa e da qualidade do prompt. Veja também prompts few-shot.

aplicação zumbi

Uma aplicação que tem um uso médio de CPU e memória inferior a 5%. Em um projeto de migração, é comum retirar essas aplicações.

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.