Escalando a infraestrutura do Amazon EKS para otimizar a computação, as cargas de trabalho e o desempenho da rede

AWS Orientação prescritiva

AWS Orientação prescritiva: Escalando a infraestrutura do Amazon EKS para otimizar a computação, as cargas de trabalho e o desempenho da rede

Table of Contents

Escalando a infraestrutura do Amazon EKS para otimizar a computação, as cargas de trabalho e o desempenho da rede

Aniket Dekate, Aniket Kurzadkar e Ishwar Chauthaiwale, da Amazon Web Services ()AWS

Novembro de 2024 (histórico do documento)

O Amazon Elastic Kubernetes Service (Amazon EKS) é um serviço gerenciado de Kubernetes. Com o Amazon EKS, você pode executar pods do Kubernetes em um ambiente de nuvem em contêineres sem precisar instalar e operar seu próprio plano de controle. Com o AWS gerenciamento do plano de controle, o Amazon EKS reduz o gerenciamento operacional organizacional. Outros benefícios do uso do Amazon EKS incluem escalabilidade, confiabilidade e segurança no ambiente de nuvem.

Este guia foi criado para ajudar as organizações a otimizar sua infraestrutura do Amazon EKS nas seguintes áreas:

- O escalonamento computacional é um componente essencial para o desempenho do aplicativo em um ambiente dinâmico do Kubernetes:
  - Alocação eficiente de recursos — Aprenda sobre técnicas para alocar recursos computados dinamicamente para atender à demanda variável.
  - Ferramentas de automação — tenha uma visão geral das ferramentas e serviços que automatizam o escalonamento computacional, reduzindo a necessidade de intervenção manual.
- O escalonamento da carga de trabalho ajuda a garantir que os aplicativos possam lidar com cargas de trabalho variadas sem degradação do desempenho:
  - Autoescalador horizontal de pods — veja detalhadamente como um HPA ajuda a escalar cargas de trabalho com base em métricas em tempo real.
  - Autoescalador proporcional de cluster — Saiba como o CPA escala e mantém automaticamente uma relação proporcional entre nós e réplicas, aumentando ou diminuindo as cargas de trabalho à medida que o tamanho do cluster muda.
  - Escalabilidade orientada por eventos — Analise as estratégias para escalar aplicativos em resposta a eventos ou acionadores específicos.
- O escalonamento de rede ajuda a manter a comunicação perfeita entre os serviços e o fluxo de dados eficiente em ambientes dinâmicos:

  - Plug-in Amazon VPC CNI — Saiba como o plug-in VPC CNI permite redes escaláveis nos clusters do Amazon EKS.
  - Rede personalizada - Analise o gerenciamento de endereços IP e a segregação do tráfego de rede nos clusters do Amazon EKS.
  - Delegação de prefixo - Tenha uma visão geral da simplificação do gerenciamento de IP em clusters grandes e escaláveis do Amazon EKS.
  - Amazon VPC Lattice — Tenha uma visão geral de como o VPC Lattice pode gerenciar várias VPCs e redes para uma escalabilidade perfeita. service-to-service
- A otimização de custos ajuda as empresas a ver onde seus recursos estão sendo gastos e a atribuir adequadamente as despesas aos departamentos ou projetos:
  - Dimensionamento correto dos recursos — considere técnicas para dimensionar os recursos da nuvem de forma adequada à carga de trabalho.
  - Monitoramento e controle de custos — Analise as ferramentas e as melhores práticas para rastrear e otimizar as despesas com a nuvem.

Cada seção se concentra em metas específicas que são necessárias para criar um ambiente de nuvem confiável, eficaz e acessível.

Objetivos

Este guia pode ajudar você e sua organização a alcançar os seguintes objetivos comerciais:

- Maior eficiência de recursos — obtenha a utilização ideal dos recursos escalando dinamicamente os recursos de computação, cargas de trabalho e rede com base nas demandas em tempo real.

  Esse objetivo enfatiza a importância de aumentar e diminuir os recursos em resposta aos padrões reais de uso. Ferramentas como autoescaladores de pod horizontais e o plug-in CNI do Amazon VPC ajudam as organizações a usar apenas os recursos de que precisam, minimizando o desperdício e maximizando o desempenho.

- Melhor desempenho do aplicativo — mantenha o alto desempenho e a capacidade de resposta dos aplicativos, mesmo sob cargas de trabalho e padrões de tráfego flutuantes.

  Esse objetivo se concentra em estratégias para ajudar a garantir que os aplicativos possam lidar com picos de tráfego e cargas de trabalho pesadas sem comprometer o desempenho.

Técnicas como escalabilidade de carga de trabalho orientada por eventos, alocação eficiente de computação e arquiteturas de rede escaláveis são fundamentais para alcançar esse objetivo.

- Escalabilidade perfeita — Permita o escalonamento suave dos componentes da infraestrutura, permitindo o crescimento e a adaptação sem esforço às mudanças nas necessidades dos negócios.

  A escalabilidade perfeita é crucial para organizações que prevêem crescimento ou experimentam níveis de tráfego variáveis. Esse objetivo aborda a importância de implementar soluções escaláveis em recursos de computação, carga de trabalho e rede, para que o dimensionamento possa ser automático, eficiente e transparente.

- Otimização de custos — Minimize os custos da nuvem enquanto mantém ou melhora o desempenho e a escalabilidade.

  A otimização de custos pode incluir a redução de despesas, como o dimensionamento correto dos recursos, o uso de soluções de escalabilidade econômicas e o monitoramento de gastos. O objetivo é equilibrar a economia de custos com a necessidade de alto desempenho e escalabilidade.

Escalabilidade computacional

O escalonamento computacional é um componente essencial para o desempenho do aplicativo em um ambiente dinâmico do Kubernetes. O Kubernetes reduz o desperdício por meio do ajuste dinâmico dos recursos de computação (como CPU e memória) em resposta à demanda em tempo real. Esse recurso ajuda a evitar o provisionamento excessivo ou insuficiente, o que também pode economizar despesas operacionais. O Kubernetes elimina efetivamente a necessidade de intervenção manual, permitindo que a infraestrutura aumente automaticamente durante os horários de pico e diminua nos períodos fora do pico.

O escalonamento computacional geral do Kubernetes automatiza o processo de escalabilidade, o que aumenta a flexibilidade e a escalabilidade do aplicativo e aprimora seu comportamento tolerante a falhas. Em última análise, os recursos do Kubernetes aprimoram a excelência operacional e a produtividade.

Esta seção aborda os seguintes tipos de escalabilidade computacional:

- Autoscaler do cluster
- Autoescalador de cluster com provisionamento excessivo
- Karpenter

Cluster AutoScaler

Dependendo das necessidades dos pods, a ferramenta Cluster Autoscaler modifica automaticamente o tamanho adicionando nós quando necessário ou removendo nós quando não são necessários e estão subutilizados.

Considere a ferramenta Cluster Autoscaler como uma solução de escalabilidade para cargas de trabalho em que a demanda aumenta gradualmente e a latência no escalonamento não é um grande problema.

A ferramenta Cluster Autoscaler fornece os seguintes recursos principais:

- Dimensionamento — Aumenta e diminui os nós dinamicamente em resposta às demandas reais de recursos.
- Agendamento de pods — ajuda a garantir que cada pod esteja operando e tenha os recursos necessários para funcionar, evitando a escassez de recursos.

- Custo-benefício — elimina as despesas desnecessárias de operar nós subutilizados ao eliminá-los.

Autoescalador de cluster com provisionamento excessivo

O autoescalador de cluster com superprovisionamento funciona de forma semelhante ao autoescalador de cluster, pois implanta nós de forma eficiente e economiza tempo executando pods de baixa prioridade nos nós. Com essa técnica, o tráfego é redirecionado para esses pods em resposta a picos repentinos na demanda, permitindo que o aplicativo continue operando sem interrupção.

O escalador automático de cluster com provisionamento excessivo oferece os recursos de pods fictícios que podem ser usados para implantar e executar nós com facilidade quando a carga de trabalho é muito grande, a latência não é necessária e o escalonamento precisa ser rápido.

O escalador automático de cluster com provisionamento excessivo fornece os seguintes recursos principais:

- Melhor capacidade de resposta — ao tornar o excesso de capacidade constantemente acessível, leva menos tempo para escalar o cluster em resposta aos picos de demanda.
- Reserva de recursos — O gerenciamento de picos inesperados no tráfego auxilia efetivamente no gerenciamento correto com pouco tempo de inatividade.
- Escalabilidade suave — Minimizar os atrasos na alocação de recursos facilita um processo de escalabilidade mais contínuo.

Karpenter

O Karpenter for Kubernetes supera a ferramenta tradicional de escalonamento automático de cluster em termos de código aberto, desempenho e personalização. Com o Karpenter, você pode iniciar automaticamente somente os recursos computacionais necessários para lidar com as demandas do seu cluster em tempo real. O Karpenter foi projetado para oferecer um dimensionamento mais eficiente e responsivo.

Aplicativos com cargas de trabalho extremamente variáveis ou complexas, nas quais decisões rápidas de escalonamento são essenciais, se beneficiam muito do uso do Karpenter. Ele se integra AWS, oferecendo melhor implantação e otimização da seleção de nós.

O Karpenter inclui os seguintes recursos principais:

- Provisionamento dinâmico — O Karpenter fornece as instâncias e os tamanhos certos para a finalidade e provisiona novos nós dinamicamente com base nos requisitos específicos dos pods.
- Programação avançada — Usando o posicionamento inteligente do pod, o Karpenter organiza os nós de forma que recursos como GPU, CPU, memória e armazenamento sejam usados da forma mais eficaz possível.
- Escalonamento rápido — O Karpenter pode escalar rapidamente, reagindo frequentemente em segundos. Essa capacidade de resposta é útil para padrões de tráfego repentino ou quando a carga de trabalho exige escalabilidade imediata.
- Eficiência de custos — Ao escolher cuidadosamente a instância mais eficaz, você pode reduzir os custos operacionais e aproveitar as alternativas adicionais de economia de custos oferecidas por AWS, como instâncias sob demanda, instâncias spot e instâncias reservadas.

Dimensionamento da carga de trabalho

O escalonamento da carga de trabalho no Kubernetes é essencial para manter o desempenho do aplicativo e a eficiência dos recursos em ambientes dinâmicos. O escalonamento ajuda a garantir que os aplicativos possam lidar com cargas de trabalho variadas sem degradação do desempenho. O Kubernetes fornece a capacidade de aumentar ou reduzir automaticamente os recursos com base em métricas em tempo real, permitindo que as organizações respondam rapidamente às mudanças no tráfego. Essa elasticidade não apenas melhora a experiência do usuário, mas também otimiza a utilização dos recursos, ajudando a minimizar os custos associados a recursos subutilizados ou superprovisionados.

Além disso, o escalonamento eficaz da carga de trabalho oferece suporte à alta disponibilidade, garantindo que os aplicativos permaneçam responsivos mesmo durante os períodos de pico de demanda. O escalonamento da carga de trabalho no Kubernetes permite que as organizações façam melhor uso dos recursos da nuvem ajustando dinamicamente a capacidade para atender às necessidades atuais.

Esta seção discute os seguintes tipos de escalabilidade da carga de trabalho:

- Autoescalador horizontal de cápsulas
- Autoescalador proporcional de cluster
- Autoescalador baseado em eventos baseado em Kubernetes

Horizontal Pod Autoscaler

O Horizontal Pod Autoscaler (HPA) é um recurso do Kubernetes que ajusta automaticamente o número de réplicas de pod em uma implantação, controlador de replicação ou conjunto com estado, com base na utilização observada da CPU ou em outras métricas selecionadas. O HPA garante que os aplicativos possam gerenciar os níveis flutuantes de tráfego e carga de trabalho sem a necessidade de intervenção manual. O HPA oferece um meio de preservar o desempenho ideal e, ao mesmo tempo, fazer uso eficaz dos recursos disponíveis.

Em contextos em que a demanda do usuário pode flutuar consideravelmente ao longo do tempo, como aplicativos da web, microsserviços e APIs, o HPA é especialmente útil.

O escalador automático Horizontal Pod fornece os seguintes recursos principais:

- Escalabilidade automática — o HPA aumenta ou diminui automaticamente o número de réplicas de pods em resposta às métricas em tempo real, garantindo que os aplicativos possam ser escalados para atender à demanda do usuário.
- Decisões baseadas em métricas — Por padrão, o HPA é dimensionado com base na utilização da CPU. No entanto, ele também pode usar métricas personalizadas, como uso de memória ou métricas específicas do aplicativo, permitindo estratégias de escalabilidade mais personalizadas.
- Parâmetros configuráveis — você pode escolher as contagens mínima e máxima de réplicas e as porcentagens de utilização desejadas, o que lhe dá autoridade sobre a gravidade da escalabilidade.
- Integração com o Kubernetes — Para monitorar e modificar recursos, o HPA trabalha em conjunto com outros elementos do ecossistema Kubernetes, incluindo o Metrics Server, a API do Kubernetes e adaptadores de métricas personalizados.
- Melhor utilização dos recursos — a HPA ajuda a garantir que os recursos sejam usados de forma eficaz, reduzindo os custos e melhorando o desempenho, modificando dinamicamente o número de pods.

Autoescalador proporcional de cluster

O Cluster Proportional Autoscaler (CPA) é um componente do Kubernetes projetado para ajustar automaticamente o número de réplicas de pods em um cluster com base no número de nós disponíveis. Ao contrário dos autoescaladores tradicionais, que escalam com base em métricas de utilização de recursos (como CPU e memória), o CPA dimensiona as cargas de trabalho em proporção ao tamanho do próprio cluster.

Essa abordagem é particularmente útil para aplicativos que precisam manter um certo nível de redundância ou disponibilidade em relação ao tamanho do cluster, como CoreDNS e outros serviços de infraestrutura. Alguns dos principais casos de uso do CPA incluem o seguinte:

- Provisionamento excessivo
- Expanda os principais serviços da plataforma
- Expanda as cargas de trabalho porque o CPA não exige um servidor de métricas ou um adaptador Prometheus

Ao automatizar o processo de escalabilidade, o CPA ajuda as empresas a manter uma distribuição equilibrada da carga de trabalho, aumentando a eficiência dos recursos e garantindo que os aplicativos sejam adequadamente provisionados para atender à demanda do usuário.

O autoescalador proporcional de cluster fornece os seguintes recursos principais:

- Escalabilidade baseada em nós — o CPA dimensiona as réplicas de acordo com o número de nós do cluster que podem ser programados, permitindo que os aplicativos se expandam ou se contraiam proporcionalmente ao tamanho do cluster.
- Ajuste proporcional — Para garantir que o aplicativo possa ser escalado de acordo com as mudanças no tamanho do cluster, o autoescalador estabelece uma relação proporcional entre o número de nós e o número de réplicas. Esse relacionamento é usado para calcular o número desejado de réplicas para uma carga de trabalho.
- Integração com componentes do Kubernetes — O CPA funciona com componentes padrão do Kubernetes, como o Horizontal Pod Autoscaler (HPA), mas se concentra especificamente na contagem de nós em vez das métricas de utilização de recursos. Essa integração permite uma estratégia de escalabilidade mais abrangente.
- Clientes da API Golang — Para monitorar o número de nós e seus núcleos disponíveis, o CPA usa clientes da API Golang que são executados dentro de pods e conversam com o servidor da API Kubernetes.
- Parâmetros configuráveis — Usando um`ConfigMap`, os usuários podem definir limites e parâmetros de escala que o CPA usa para modificar seu comportamento e garantir que ele siga o plano de escalabilidade pretendido.

Autoescalador baseado em eventos baseado em Kubernetes

O Event Driven Autoscaler (KEDA) baseado em Kubernetes é um projeto de código aberto que permite que as cargas de trabalho do Kubernetes sejam escaladas com base no número de eventos que precisam ser processados. O KEDA aprimora a escalabilidade dos aplicativos, permitindo que eles respondam dinamicamente a cargas de trabalho variáveis, especialmente aquelas que são orientadas por eventos.

Ao automatizar o processo de escalabilidade com base em eventos, a KEDA ajuda as organizações a otimizar a utilização de recursos, melhorar o desempenho do aplicativo e reduzir os custos associados ao provisionamento excessivo. Essa abordagem é especialmente valiosa para aplicativos

que experimentam padrões de tráfego variados, como microsserviços, funções sem servidor e sistemas de processamento de dados em tempo real.

O KEDA fornece os seguintes recursos principais:

- Escalonamento orientado por eventos — O KEDA permite que você defina regras de escalonamento com base em fontes externas de eventos, como filas de mensagens, solicitações HTTP ou métricas personalizadas. Esse recurso ajuda a garantir que os aplicativos sejam dimensionados em resposta à demanda em tempo real.
- Componente leve — O KEDA é um componente leve e de propósito único que não requer muita configuração ou sobrecarga para ser facilmente integrado aos clusters Kubernetes existentes.
- Integração com o Kubernetes — A KEDA amplia os recursos dos componentes nativos do Kubernetes, como o Horizontal Pod Autoscaler (HPA). A KEDA adiciona recursos de escalonamento orientados por eventos a esses componentes, aprimorando-os em vez de substituí-los.
- Support para várias fontes de eventos — O KEDA é compatível com uma ampla variedade de fontes de eventos, incluindo plataformas de mensagens populares como RabbitMQ, Apache Kafka e outras. Devido a essa adaptabilidade, você pode personalizar o escalonamento para se adequar à sua arquitetura exclusiva orientada a eventos.
- Escaladores personalizados — Usando escaladores personalizados, você pode designar métricas específicas que a KEDA pode usar para iniciar ações de escalabilidade em resposta a requisitos ou lógicas comerciais específicas.
- Configuração declarativa — De acordo com os princípios do Kubernetes, você pode usar o KEDA para descrever o comportamento de escalabilidade de forma declarativa usando recursos personalizados do Kubernetes para definir como o escalonamento deve acontecer.

Escalabilidade de rede

O escalonamento da rede no Kubernetes é fundamental para manter a comunicação perfeita entre os serviços e dar suporte ao fluxo de dados eficiente em ambientes dinâmicos. A escalabilidade da infraestrutura de rede ajuda a garantir que o cluster possa lidar com vários níveis de tráfego sem enfrentar gargalos ou problemas de latência. O Kubernetes fornece ferramentas e mecanismos para escalar os recursos da rede, permitindo que as organizações mantenham um desempenho ideal à medida que os padrões de tráfego mudam.

Essa elasticidade no escalonamento da rede aprimora a experiência geral do usuário, garantindo conexões rápidas e confiáveis. O escalonamento da rede também otimiza o uso dos recursos da rede, ajudando a reduzir os custos associados a componentes de rede subutilizados ou sobrecarregados.

Além disso, o escalonamento eficaz da rede é vital para oferecer suporte à alta disponibilidade e resiliência. Ao ajustar dinamicamente a capacidade e o roteamento da rede, as organizações podem garantir que os serviços permaneçam acessíveis e responsivos mesmo durante períodos de pico de demanda ou picos de tráfego inesperados. Essa abordagem permite uma melhor utilização dos recursos de rede em nuvem, garantindo que a infraestrutura esteja sempre alinhada aos requisitos atuais.

Esta seção aborda os seguintes tipos de escalabilidade de rede:

- Plug-in CNI da Amazon VPC para Kubernetes
- Rede personalizada
- Delegação de prefixo
- Amazon VPC Lattice

Plug-in CNI da Amazon VPC para Kubernetes

O plug-in Amazon VPC Container Network Interface (CNI) para Kubernetes é um componente essencial no Amazon EKS. O plug-in VPC CNI fornece recursos avançados de rede ao integrar os pods do Kubernetes com o Amazon VPC. Com esse plug-in, cada pod recebe um endereço IP exclusivo da nuvem privada virtual (VPC), aprimorando assim o isolamento e o desempenho da rede. À medida que os clusters crescem e as demandas de rede flutuam, o plug-in CNI da Amazon VPC desempenha um papel fundamental na garantia de operações de rede eficientes e escaláveis.

O plug-in gerencia automaticamente a alocação e o roteamento de endereços IP na VPC, simplificando o gerenciamento da rede e reduzindo o risco de conflitos de IP. Ele oferece suporte a recursos como delegação de prefixos, o que permite um gerenciamento de IP mais flexível.

O plug-in VPC CNI ajuda as organizações a otimizar o desempenho da rede, aprimorar a segurança e reduzir o risco de esgotamento do IP. Esses recursos são especialmente valiosos para ambientes dinâmicos de grande escala em que as demandas de rede flutuam, como arquiteturas de microsserviços, cargas de trabalho de alta densidade e aplicativos multilocatários.

O plug-in CNI da Amazon VPC fornece os seguintes recursos principais:

- Rede aprimorada — O plug-in VPC CNI permite que cada pod receba seu próprio endereço IP diretamente da VPC, fornecendo isolamento e desempenho de rede robustos. Essa abordagem é crucial para cargas de trabalho que exigem alta taxa de transferência de rede e baixa latência.
- Delegação de prefixo — Para superar problemas de exaustão de endereços IP em grandes clusters, a delegação de prefixos aloca dinamicamente blocos maiores de dois nós, que são então IPs subdivididos para uso do pod. Essa abordagem garante a utilização eficiente do IP e simplifica o dimensionamento da rede.
- Rede personalizada — Os usuários podem configurar interfaces de rede personalizadas (ENIs) para pods, o que ajuda a distribuir o tráfego de pod em várias interfaces, reduzindo o congestionamento da rede e melhorando a escalabilidade.
- Support for IPv6 — Ao habilitar IPv6 clusters do Amazon EKS, os usuários podem expandir significativamente o espaço de endereço IP disponível, facilitando a escalabilidade de aplicativos grandes e distribuídos sem as restrições ou limitações. IPv4
- Integração com o Kubernetes — O plug-in VPC CNI funciona perfeitamente com os componentes de rede do Kubernetes, garantindo que IPs sejam gerenciados de forma eficiente em pods, serviços e endpoints externos, além de oferecer suporte a recursos avançados, como grupos de segurança para pods.

Redes personalizadas

A rede personalizada no Amazon EKS permite a atribuição de interfaces de rede específicas a pods, fornecendo controle aprimorado sobre o gerenciamento de endereços IP e o tráfego de rede. Essa abordagem é especialmente útil em cenários em que o esgotamento do endereço IP é uma preocupação ou quando há a necessidade de segregar o tráfego de rede por motivos de segurança, conformidade ou desempenho. A rede personalizada ajuda as organizações a gerenciar

com eficiência o espaço de endereços IP, segregar o tráfego e garantir um desempenho de rede escalável.

Com a rede personalizada, os administradores podem gerenciar os recursos da rede com mais eficiência. Os administradores podem usar redes personalizadas para ajudar a garantir que os pods tenham o isolamento de rede necessário e que o cluster possa ser escalado sem se deparar com limitações de endereço IP.

A rede personalizada fornece os seguintes recursos principais:

- Gerenciamento aprimorado de IP — A rede personalizada permite a atribuição de interfaces de rede específicas (ENIs) aos pods, ajudando a gerenciar o esgotamento do endereço IP distribuindo o tráfego de pods em vários. ENIs Esse recurso é particularmente importante em clusters com cargas de trabalho de alta densidade.
- Segregação de tráfego — Com interfaces de rede personalizadas, você pode separar o tráfego do pod com base em critérios específicos, como tipo de aplicativo ou requisitos de segurança. Essa abordagem fornece maior controle sobre como o tráfego flui dentro e fora do cluster.
- Support for IPv6 — A rede personalizada no Amazon EKS também oferece suporte IPv6, oferecendo uma solução para as limitações de IPv4 endereços. A rede pode ser escalada de forma eficiente sem conflitos de endereço IP, mesmo em implantações em grande escala.
- Escalabilidade e flexibilidade — À medida que o cluster se expande, a rede personalizada permite o gerenciamento dinâmico das interfaces de rede. Os novos pods recebem recursos de rede apropriados sem intervenção manual. Essa abordagem ajuda a manter um ambiente de rede flexível e escalável que pode se adaptar às mudanças nas cargas de trabalho.

Delegação de prefixo

A delegação de prefixos no Kubernetes, especialmente no Amazon EKS, foi projetada para simplificar e otimizar o gerenciamento de endereços IP à medida que os clusters se expandem. Ao alocar dinamicamente blocos maiores de endereços IP (prefixos) aos nós, a delegação de prefixos reduz o risco de esgotamento do IP e simplifica o gerenciamento do espaço IP.

Essa abordagem aumenta a eficiência da rede, minimiza a fragmentação e ajuda os clusters a escalarem sem problemas, sem ajustes manuais no intervalo de IP. A delegação de prefixos é particularmente valiosa para implantações em grande escala, cargas de trabalho de alta densidade e ambientes em que o gerenciamento flexível e dinâmico de IP é essencial para manter o desempenho e a escalabilidade da rede.

A delegação de prefixos fornece os seguintes recursos principais:

- Gerenciamento eficiente de endereços IP — a delegação de prefixos permite a alocação dinâmica de intervalos de IP, reduzindo o risco de exaustão de IP e garantindo o uso eficiente do espaço IP disponível.
- Gerenciamento de rede simplificado — ao permitir que os nós lidem com suas próprias alocações de IP, a delegação de prefixos minimiza a fragmentação da rede e simplifica o processo de roteamento, facilitando a escalabilidade dos clusters conforme necessário.
- Support para implantações em grande escala — Em grandes clusters com cargas de trabalho de alta densidade, a delegação de prefixos permite um escalonamento contínuo, permitindo que novos nós se juntem ao cluster sem ajustes manuais no intervalo de IP.

Amazon VPC Lattice

O Amazon VPC Lattice permite uma service-to-service comunicação eficiente e segura dentro e fora dela VPCs, especialmente em arquiteturas de microsserviços. O VPC Lattice usa medidas de segurança, como grupos de segurança e listas de controle de acesso à rede (rede ACLs), além da integração AWS Identity and Access Management (IAM) para autenticação refinada de aplicativos. Um serviço de proxy de camada 7 no centro do VPC Lattice oferece conexão, balanceamento de carga, autenticação, autorização, observabilidade, gerenciamento de tráfego e descoberta de serviços.

Ao simplificar as configurações de rede e segurança, o VPC Lattice ajuda as organizações a otimizar o gerenciamento de tráfego, aprimorar o desempenho do aplicativo e escalar perfeitamente entre vários e. VPCs Regiões da AWS Isso é especialmente valioso para aplicativos distribuídos que exigem redes consistentes e confiáveis, como microsserviços, implantações entre regiões e ambientes complexos nativos da nuvem.

O Amazon VPC Lattice fornece os seguintes recursos principais:

- Service-to-service rede — O VPC Lattice simplifica a configuração de rede e segurança entre serviços em uma arquitetura de microsserviços. Ele fornece uma plataforma unificada para gerenciar a comunicação, para que os serviços possam ser escalados de forma independente, mantendo o alto desempenho e a segurança.
- Rede entre VPC — o VPC Lattice é crucial para gerenciar o tráfego em várias regiões. VPCs Ele fornece uma estrutura de rede consistente que permite que os serviços se comuniquem sem problemas, independentemente de sua localização física. Esse recurso é particularmente

importante para aplicativos de grande escala que abrangem várias regiões VPCs ou regiões geográficas.

- Gerenciamento de segurança aprimorado — Ao integrar políticas de segurança diretamente na camada de rede, o VPC Lattice service-to-service oferece suporte a uma comunicação segura e eficiente. Esse recurso reduz a complexidade do gerenciamento da segurança em um ambiente distribuído, facilitando o dimensionamento e reduzindo a sobrecarga operacional.
- Gerenciamento de tráfego simplificado — o VPC Lattice oferece recursos avançados de gerenciamento de tráfego, incluindo roteamento, balanceamento de carga e mecanismos de failover. Com esses recursos, o tráfego é distribuído de forma eficiente entre os serviços, otimizando o desempenho da rede e aprimorando a escalabilidade do aplicativo.

Otimização de custo

Para apoiar o controle efetivo de recursos, a minimização de custos do Kubernetes é crucial para empresas que usam essa tecnologia de orquestração de contêineres. É difícil monitorar adequadamente os gastos nas configurações do Kubernetes devido à sua complexidade, que inclui vários componentes, como pods e nós. Por meio da aplicação de técnicas de otimização de custos, as empresas podem ver onde seus recursos estão sendo gastos e atribuir adequadamente as despesas aos departamentos ou projetos.

Embora o escalonamento dinâmico tenha vantagens, se não for gerenciado adequadamente, pode resultar em despesas imprevistas. O gerenciamento eficiente de custos ajuda a alocar recursos somente quando eles são realmente necessários, evitando aumentos imprevistos nas despesas.

Esta seção discute as seguintes abordagens para otimização de custos:

- Cubecost
- Cachinhos Dourados
- AWS Fargate
- Instâncias spot
- Instâncias reservadas
- AWS Instâncias de Graviton

Kubecost

O Kubecost é uma solução de gerenciamento de custos que ajuda as empresas a rastrear, controlar e maximizar seus gastos com infraestrutura em nuvem. Ele foi feito especificamente para clusters Kubernetes. O Kubecost fornece informações sobre a utilização de recursos e o reconhecimento de custos em tempo real, permitindo que você entenda melhor onde e quanto dos seus recursos de nuvem estão sendo usados. Com esses insights, você pode otimizar seus gastos com infraestrutura, melhorar a eficiência dos recursos e tomar decisões mais informadas sobre seus investimentos em nuvem.

O Kubecost fornece os seguintes recursos principais:

- Alocação de custos — O Kubecost oferece uma alocação completa de custos para os recursos do Kubernetes, incluindo cargas de trabalho, serviços, namespaces e rótulos. Esse recurso ajuda as equipes a monitorar os custos por ambiente, projeto ou equipe.
- Monitoramento de custos em tempo real — Ele oferece monitoramento em tempo real dos custos da nuvem, fornecendo às organizações informações imediatas sobre os padrões de gastos e ajudando a evitar custos excessivos inesperados.
- Recomendações de otimização — O Kubecost oferece sugestões práticas para minimizar a utilização de recursos, incluindo a redução de recursos ociosos, o dimensionamento correto das cargas de trabalho e a maximização das despesas de armazenamento.
- Orçamento e alertas — Os usuários do Kubecost podem criar orçamentos e receber lembretes quando uma despesa se aproxima ou ultrapassa os critérios predeterminados. Esse recurso ajuda as equipes a cumprir as restrições financeiras.

Cachinhos Dourados

O Goldilocks é um utilitário do Kubernetes projetado para ajudar os usuários a otimizar suas solicitações de recursos e limites para cargas de trabalho do Kubernetes. Ele fornece recomendações sobre como configurar os recursos de CPU e memória para contêineres em execução em um cluster Kubernetes. Essas recomendações ajudam você a garantir que os aplicativos tenham o número certo de recursos para um desempenho eficiente sem desperdício. Essa otimização pode levar à economia de custos, melhor desempenho e uso mais eficiente dos clusters Kubernetes.

Goldilocks fornece os seguintes recursos principais:

- Recomendações de recursos — Goldilocks determina as configurações ideais para solicitações e restrições de recursos analisando estatísticas anteriores de consumo de CPU e memória para cargas de trabalho do Kubernetes. Ao fazer isso, fica mais fácil evitar o provisionamento insuficiente ou excessivo, o que pode resultar em problemas de desempenho e desperdício de recursos.
- Integração com VPA — A Goldilocks utiliza o Kubernetes Vertical Pod Autoscaler (VPA) para coletar dados e fornecer recomendações. Ele é executado em um “modo de recomendação”, o que significa que, na verdade, não altera as configurações dos recursos, mas oferece orientação sobre quais devem ser essas configurações.

- Análise baseada em namespaces — O Goldilocks oferece a capacidade de regular com precisão quais cargas de trabalho são otimizadas e monitoradas, permitindo que você direcione namespaces específicos para análise.
- Painel visual — O painel baseado na web exibe visualmente as solicitações e restrições de recursos sugeridas, o que facilita a compreensão e a ação dos dados.
- Operação não intrusiva — Goldilocks não altera a configuração do cluster porque opera no modo de recomendação. Se quiser, você pode aplicar manualmente as configurações de recursos recomendadas depois de analisar as recomendações.

AWS Fargate

No contexto do Amazon EKS, https://docs.aws.amazon.com/eks/latest/userguide/fargate.htmlAWS Fargate permite que você execute pods do Kubernetes sem gerenciar as instâncias subjacentes da Amazon. EC2 É um mecanismo de computação sem servidor que permite que você se concentre na implantação e na escalabilidade de aplicativos em contêineres sem se preocupar com a infraestrutura.

AWS Fargate fornece os seguintes recursos principais:

- Sem gerenciamento de infraestrutura — o Fargate elimina a necessidade de provisionar, gerenciar ou escalar EC2 instâncias da Amazon ou nós do Kubernetes. AWS lida com todo o gerenciamento da infraestrutura, incluindo patches e escalabilidade.
- Isolamento em nível de pod — diferentemente dos nós de trabalho baseados na Amazon, o EC2 Fargate fornece isolamento em nível de tarefa ou de pod. Cada pod é executado em seu próprio ambiente computacional isolado, o que aumenta a segurança e o desempenho.
- Escalabilidade automática — O Fargate escala automaticamente os pods do Kubernetes com base na demanda. Você não precisa gerenciar políticas de escalabilidade ou pools de nós.
- Cobrança por segundo — você paga apenas pelos recursos de vCPU e memória consumidos por cada pod pela duração exata em que ele é executado, o que é uma opção econômica para determinadas cargas de trabalho.
- Redução da sobrecarga — Ao eliminar a necessidade de gerenciar EC2 instâncias, o Fargate permite que você se concentre na criação e no gerenciamento de seus aplicativos, em vez de nas operações de infraestrutura.

Instâncias spot

As instâncias spot oferecem economias significativas em relação aos preços de instâncias sob demanda e são uma opção acessível para executar os nós de EC2 trabalho da Amazon em um cluster do Amazon EKS. No entanto, AWS pode interromper as instâncias spot caso seja necessária a capacidade da instância sob demanda. AWS podem recuperar instâncias spot com um aviso prévio de 2 minutos quando a capacidade é necessária, tornando-as menos confiáveis para cargas de trabalho críticas e monitoradas.

Para cargas de trabalho sensíveis ao custo e capazes de suportar interrupções, as instâncias spot no Amazon EKS são uma boa opção. Usar uma combinação de instâncias spot e instâncias sob demanda em um cluster Kubernetes ajuda você a economizar dinheiro sem sacrificar a disponibilidade de cargas de trabalho vitais.

As instâncias spot fornecem os seguintes recursos principais:

- Economia de custos — As instâncias spot podem ser mais baratas do que os preços das instâncias sob demanda, o que as torna ideais para cargas de trabalho econômicas.
- Ideal para cargas de trabalho tolerantes a falhas — Adequado para cargas de trabalho sem estado e tolerantes a falhas, como processamento em lote, tarefas de CI/CD, aprendizado de máquina ou processamento de dados em grande escala, em que as instâncias podem ser substituídas sem grandes interrupções.
- Integração de grupos com escalabilidade automática — O Amazon EKS integra instâncias spot com o Kubernetes Cluster Autoscaler, que pode substituir automaticamente os nós de instância spot interrompidos por outras instâncias spot ou instâncias sob demanda disponíveis.

Instâncias reservadas

No Amazon EKS, as Instâncias Reservadas são um modelo de preços para os nós de EC2 trabalho da Amazon que executam suas cargas de trabalho do Kubernetes. Ao usar instâncias reservadas, você se compromete a usar tipos específicos de instância por um período de 1 ou 3 anos, em troca da economia de custos em comparação com os preços das instâncias sob demanda. Reservar instâncias no Amazon EKS é uma forma acessível de realizar cargas de trabalho consistentes e de longo prazo nos nós de EC2 trabalho da Amazon.

As instâncias reservadas são comumente usadas pela Amazon EC2. No entanto, os nós de trabalho em seu cluster Amazon EKS (que são EC2 instâncias) também podem se beneficiar desse modelo econômico, desde que a carga de trabalho exija um uso previsível e de longo prazo.

Serviços de produção, bancos de dados e outros aplicativos monitorados que precisam de alta disponibilidade e desempenho consistente são exemplos de cargas de trabalho estáveis que são adequadas para instâncias reservadas.

As instâncias reservadas fornecem os seguintes recursos principais:

- Economia de custos — As instâncias reservadas oferecem economia em comparação com as instâncias sob demanda, dependendo da duração do prazo (1 ou 3 anos) e do plano de pagamento (pagamento adiantado total, adiantado parcial ou sem pagamento adiantado).
- Compromisso de longo prazo — Você se compromete com um prazo de 1 ou 3 anos para um tipo, tamanho e tamanho de instância específicos. Região da AWS Isso é ideal para cargas de trabalho estáveis e executadas continuamente ao longo do tempo.
- Preços previsíveis — Como você está comprometido com um período específico, as Instâncias Reservadas oferecem custos mensais ou iniciais previsíveis, facilitando o orçamento para cargas de trabalho de longo prazo.
- Flexibilidade de instância — Com as instâncias reservadas conversíveis, você pode alterar o tipo, a família ou o tamanho da instância durante o período da reserva. As instâncias reservadas conversíveis oferecem mais flexibilidade do que as instâncias reservadas padrão, que não permitem alterações.
- Capacidade garantida — As instâncias reservadas garantem que a capacidade esteja disponível na zona de disponibilidade em que a reserva é feita, o que é crucial para cargas de trabalho críticas que precisam de potência computacional consistente.
- Sem risco de interrupção — Diferentemente das Instâncias Spot, as Instâncias Reservadas não estão sujeitas à interrupção por. AWS Isso os torna ideais para executar cargas de trabalho de missão crítica que exigem tempo de atividade garantido.

AWS Instâncias de Graviton

AWS O Graviton é uma família de processadores baseados em ARM projetados AWS para fornecer melhor desempenho e economia para cargas de trabalho na nuvem. No contexto do Amazon EKS, você pode usar instâncias do Graviton como nós de trabalho para executar suas cargas de trabalho do Kubernetes, oferecendo ganhos significativos de desempenho e economia de custos.

As instâncias Graviton são uma excelente opção para aplicativos nativos da nuvem e com uso intensivo de computação, pois oferecem uma relação preço-desempenho mais alta do que as instâncias x86. No entanto, ao considerar a adoção de instâncias do Graviton, leve em consideração a compatibilidade com o ARM.

AWS As instâncias do Graviton oferecem os seguintes recursos principais:

- Arquitetura baseada em ARM — Os processadores AWS Graviton são baseados na arquitetura ARM, que é diferente das arquiteturas x86 tradicionais, mas altamente eficiente para muitas cargas de trabalho.
- Econômico — as EC2 instâncias da Amazon baseadas no Graviton normalmente oferecem melhor relação preço-desempenho em comparação às instâncias baseadas em x86. EC2 Isso os torna uma opção atraente para clusters Kubernetes que executam o Amazon EKS.
- Desempenho — Os processadores Graviton2, a segunda geração do AWS Graviton, oferecem melhorias significativas em termos de desempenho computacional, taxa de transferência de memória e eficiência energética. Eles são ideais para cargas de trabalho com uso intenso de CPU e memória.
- Diversos tipos de instância — as instâncias do Graviton vêm em várias famílias, como t4g, m7g, c7g e r7g, abrangendo uma variedade de casos de uso, desde cargas de trabalho de uso geral até cargas de trabalho otimizadas para computação, otimizadas para memória e intermitentes.
- Grupos de nós do Amazon EKS — Você pode configurar grupos de nós gerenciados pelo Amazon EKS ou grupos de nós autogerenciados para incluir instâncias baseadas em Graviton. Com essa abordagem, você pode executar cargas de trabalho otimizadas para a arquitetura ARM no mesmo cluster Kubernetes junto com instâncias baseadas em x86.

Próximas etapas

Este guia fornece informações para ajudar você a otimizar o Amazon EKS com relação à escalabilidade computacional, escalabilidade de carga de trabalho, escalabilidade de rede e otimização de custos. Ao entender e aplicar esses conceitos, as organizações podem obter um ambiente de nuvem altamente eficiente, escalável e econômico que atenda às suas necessidades dinâmicas.

A implementação eficaz do dimensionamento da computação e da carga de trabalho ajuda a garantir que os recursos sejam usados com eficiência e que os aplicativos mantenham o alto desempenho mesmo nos horários de pico. A adoção de técnicas de escalabilidade de rede, como redes personalizadas e delegação de prefixos, oferece suporte ao gerenciamento de recursos de rede e à escalabilidade perfeita. Enfatizar a otimização de custos ajuda as organizações a equilibrar desempenho com eficiência financeira.

Integrar essa orientação à sua estratégia de nuvem pode ajudá-lo a aprimorar o desempenho e a escalabilidade da sua infraestrutura e gerar economia de custos. Essa abordagem abrangente pode permitir que você crie um ambiente de nuvem robusto que suporte o crescimento da sua organização e se adapte às demandas comerciais em constante mudança.

Recursos

AWS blogs

- Construindo para otimização de custos e resiliência para EKS com instâncias spot
- AWS Combinação do Graviton com o x86 CPUs para otimizar o custo e a resiliência usando o Amazon EKS

AWS documentação

- CNI da Amazon VPC
- Amazon Elastic Kubernetes Service AWS (whitepaper: Visão geral das opções de implantação ativadas) AWS
- Guia de melhores práticas do Amazon EKS
- Karpenter
- Saiba mais sobre o Kubecost
- Simplifique o gerenciamento da computação com AWS Fargate

Outros recursos

- Escalonamento automático de clusters (documentação do Kubernetes)
- Goldilocks: uma ferramenta de código aberto para recomendar solicitações de recursos (Fairwinds Blog)
- Escalonamento automático de pods horizontais (documentação do Kubernetes)
- Kubecost (documentação do Kubecost)
- Escalonamento automático orientado por eventos do Kubernetes (documentação da KEDA)

Histórico do documento

A tabela a seguir descreve mudanças significativas neste guia, Escalando a infraestrutura do Amazon EKS para otimizar a computação, as cargas de trabalho e o desempenho da rede. Se desejar receber notificações sobre futuras atualizações, inscreva-se em um feed RSS.

Alteração	Descrição	Data
Publicação inicial	—	11 de novembro de 2024

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:

- Refactor/re-architect — mova um aplicativo e modifique sua arquitetura aproveitando ao máximo os recursos nativos da nuvem para melhorar a agilidade, o desempenho e a escalabilidade. Isso normalmente envolve a portabilidade do sistema operacional e do banco de dados. Exemplo: migre seu banco de dados Oracle local para a Amazon PostgreSQL-Compatible Aurora Edition.
- 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: Migre 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

A2A () Agent-to-Agent

Um protocolo com estado para colaboração entre agentes, apoiando a delegação de tarefas e a transferência de estados.

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.

Agente

Um sistema de IA que pode raciocinar, planejar e realizar ações de forma autônoma usando ferramentas para atingir metas.

Agente Ops

Práticas operacionais para criar, testar, implantar e executar agentes de IA na produção em grande escala.

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 as AIOps são usadas na estratégia de migração para a AWS , 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 implantação

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 Implementar procedimentos de quebra de vidros na AWS Well-Architected orientação.

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.

Desenvolvedor cidadão

Um usuário corporativo que cria aplicativos de IA usando plataformas sem code/low código sem habilidades técnicas especializadas.

criptografia no lado do cliente

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

Centro de Excelência da 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 postagens do CCoE 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: realizar investimentos fundamentais para escalar sua adoção da nuvem (por exemplo, criar uma zona de pouso, definir um CCoE, estabelecer um modelo de operações)
- Migração: migrar aplicações individuais
- Re-invention — Otimizando produtos e serviços e inovando na nuvem

Esses estágios foram definidos por Stephen Orban na postagem do blog The Journey Toward Cloud-First & the Stages of Adoption no blog Nuvem AWS Enterprise Strategy. 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 CI/CD pipeline 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

Uma coleção 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 na AWS Well-Architected Estrutura. 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.

defesa completa

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 abordagem de defesa aprofundada 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 na AWS Well-Architected estrutura.

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 Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston: Addison-Wesley Professional, 2003). Para obter informações sobre como você pode usar o design orientado por domínio com o padrão strangler fig, consulte Modernizando os serviços web legados da Microsoft ASP.NET (ASMX) de forma incremental 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. Big-endian os sistemas armazenam primeiro o byte mais significativo. Little-endian os sistemas armazenam primeiro o byte menos significativo.

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 contextual, em que os modelos aprendem com exemplos (fotos) incorporados aos prompts. Few-shot a solicitação pode ser eficaz 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 treina em grandes conjuntos de dados generalizados e não rotulados. Os 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?.

Gateway FM

Um intermediário centralizado que controla e normaliza o acesso aos modelos de fundação. Também conhecido como gateway LLM.

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 gerenciar recursos, políticas e conformidade em todas as unidades organizacionais (UOs). 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 .

grades de proteção (IA)

Mecanismos de segurança que filtram, validam e restringem as entradas e saídas dos agentes para ajudar a garantir um comportamento de IA responsável e seguro.

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.

humano no circuito (HiTL)

Um padrão de fluxo de trabalho em que a execução do agente é pausada para análise e aprovação humana em pontos críticos de decisão.

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 típico de uma DevOps versão.

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 as melhores práticas de implantação usando infraestrutura imutável na AWS Well-Architected Estrutura.

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 segurança da AWS recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação 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 fabricação por meio de avanços na conectividade, dados em tempo real, automação, análise e. AI/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 das Coisas Industrial (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 Construir uma estratégia de transformação digital para a Internet das Coisas Industrial (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 segurança da AWS recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação 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 é grande modelo de linguagem (LLM)?.

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.

MCP

Consulte Protocolo de contexto do modelo.

Protocolo de contexto para modelos (MCP)

Um protocolo sem estado para comunicação entre agentes e ferramentas.

Servidor MCP

Um serviço que expõe uma ou mais ferramentas por meio do Model Context Protocol.

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 Criação de mecanismos na AWS Well-Architected estrutura.

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, máquina a máquina (M2M), baseado no padrão, para dispositivos de IoT com recursos publish/subscribelimitados.

microsserviço

Um serviço pequeno e independente que se comunica por meio de APIs bem definidas e normalmente pertence a equipes pequenas e autônomas. 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 APIs leves. 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

Cross-functional equipes que simplificam a migração de cargas de trabalho 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, a AWS Well-Architected Estrutura 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.

Comunicação de processo aberto - Arquitetura unificada (OPC-UA)

Um protocolo de comunicação máquina a máquina (M2M) para automação industrial. 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) na AWS Well-Architected Estrutura.

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 necessária 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 do S3 Regiões da AWS, à criptografia do lado do servidor com AWS KMS (SSE-KMS) e à dinâmica `PUT` e `DELETE` às solicitações ao bucket do S3.

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 segurança da AWS recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação 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 armazena informações sobre como você quer que o Amazon Route 53 responda a consultas ao DNS para um domínio e seus subdomínios dentro de uma 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 no AWS Organizations. As SCPs definem barreiras de proteção ou estabelecem limites para as ações que um administrador pode delegar a usuários ou perfis. É possível usar SCPs como listas de permissão ou 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.

Inteligência artificial sombria

Aplicativos de IA não autorizados criados ou usados fora dos canais controlados dentro de uma organização.

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.

modelo dividir e semear

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 Modernizando os serviços web legados da Microsoft ASP.NET (ASMX) de forma incremental 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

Key-value pares que atuam como metadados para organizar seus AWS recursos. 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.

ferramenta

Uma função ou API que um agente pode invocar para realizar operações em sistemas externos.

gateway de trânsito

Um hub de trânsito de rede que pode ser usado para interconectar as VPCs e as redes on-premises. 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 de VPC

Uma conexão entre duas VPCs que permite rotear 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.