Estratégias do Protocolo de Contexto de Modelo em AWS

AWS Orientação prescritiva

AWS Orientação prescritiva: Estratégias do Protocolo de Contexto de Modelo em AWS

Table of Contents

Estratégias do Protocolo de Contexto de Modelo em AWS

Amazon Web Services (colaboradores)

Março de 2026 (histórico do documento)

Este guia pode ajudá-lo a desenvolver e implementar estratégias do Model Context Protocol (MCP) em toda a sua organização para apoiar sua jornada de IA agente. À medida que os agentes e os modelos de linguagem se tornam cada vez mais centrais para as operações comerciais, estabelecer uma estratégia de MCP é fundamental para soluções de agentes bem-sucedidas.

Este guia explora três pilares fundamentais para criar uma estratégia de MCP: design de ferramentas MCP, hospedagem de servidores MCP e governança de MCP. Ao abordar esses componentes interconectados, as organizações podem criar sistemas escaláveis, seguros e eficazes para gerenciar o contexto do modelo em suas implementações de IA. Essa orientação fornece insights acionáveis e orientação estratégica para organizações em qualquer estágio da jornada de IA de uma organização, desde a experimentação inicial até as implantações de produção em grande escala. Isso os ajuda a desenvolver soluções MCP personalizadas que se alinham às suas necessidades e objetivos específicos.

Essas melhores práticas são derivadas de implementações reais de organizações que implantam MCP em escala corporativa, de uma análise dos padrões atuais de especificação de MCP e das lições aprendidas com aplicativos personalizados de modelo de linguagem grande (LLM) em produção.

Os sistemas de IA são cada vez mais sofisticados e robustos LLMs em uma ampla variedade de casos de uso. LLMs se destacam na compreensão da linguagem natural, na geração de respostas semelhantes às humanas e no raciocínio sobre informações complexas. No entanto, para transformar interfaces LLMs de conversação em sistemas que podem realizar tarefas complexas de forma autônoma, as organizações estão adotando arquiteturas de IA agênticas, sistemas de IA que podem perceber seu ambiente, raciocinar sobre metas, tomar decisões autônomas, orquestrar várias etapas e realizar ações para alcançar objetivos em nome dos usuários. Essa abordagem agente ajuda as organizações a criar sistemas de IA que podem entender a intenção do usuário por meio da linguagem natural, coordenar de forma autônoma várias fontes de dados e ferramentas e oferecer experiências personalizadas em uma escala que não era possível com os padrões tradicionais de solicitação-resposta. Para tornar esses agentes mais capazes, as organizações precisam fornecer acesso às ferramentas e dados existentes para enriquecer a compreensão contextual do agente e permitir que ele atue em nome do usuário.

O MCP fornece um protocolo padronizado para integração de ferramentas de IA, permitindo uma comunicação consistente entre agentes e recursos externos. Embora o próprio MCP defina o padrão de comunicação, implementá-lo de forma eficaz requer uma análise cuidadosa dos padrões arquitetônicos, modelos de segurança, práticas operacionais e estratégias de otimização de desempenho para obter soluções escaláveis, seguras e sustentáveis.

Este guia sintetiza as lições aprendidas nas implantações corporativas de MCP, fornecendo recomendações práticas que se alinham ao Well-Architected Framework.AWS Ele abrange estratégias para design de ferramentas MCP, hospedagem de servidores MCP e governança de MCP, que são essenciais para criar suas próprias soluções MCP. As recomendações neste guia mapeiam os seguintes cinco pilares do AWS Well-Architected Framework:

- Segurança — isolamento de token, credenciais com escopo reduzido, autorização separada read/write
- Excelência operacional — métricas de precisão de seleção de ferramentas, conjuntos de dados dourados para testes de regressão
- Confiabilidade — Limitação da taxa por usuário e por ferramenta, redução de carga
- Eficiência de desempenho — ferramentas com escopo de fluxo de trabalho, filtragem de ferramentas, pesquisa semântica para reduzir o uso da janela de contexto
- Otimização de custos — servidores MCP reutilizáveis em todas as equipes, custos de token reduzidos por solicitação por meio da filtragem de ferramentas

Público-alvo

Este guia é destinado a arquitetos, desenvolvedores e líderes de tecnologia que estão implementando soluções de IA agênticas em suas organizações. Para entender os conceitos deste guia, você deve entender como LLMs funciona e ter conhecimento básico sobre MCP, ferramentas e engenharia rápida.

Objetivos

Criar sistemas de IA da Agentic que estejam prontos para produção significa resolver juntos a governança, a otimização e a segurança para apoiar as políticas da sua organização. A seguir, explica como esse guia aborda esses objetivos:

- Governança — Sem governança centralizada, você não pode responder a perguntas de auditoria sobre suas cargas de trabalho de IA, incluindo quais agentes acessaram quais dados, com quais

permissões e quando. Você também não pode impor o controle de versão. A seção de estratégia de hospedagem MCP deste guia explica como os usuários podem estar executando servidores MCP locais desatualizados com vulnerabilidades conhecidas devido à falta de fiscalizações sistemáticas.

Para setores regulamentados, a governança é fundamental. Os auditores querem ver a aplicação de políticas e o rastreamento do uso de ferramentas em todos os agentes em um único painel. A governança do MCP fornece isso.

Seguindo as recomendações deste guia, você pode melhorar a precisão das tarefas em 28 a 32% em benchmarks revisados por pares. Para obter mais informações, consulte MARCO: Multi-Agent Real-Time Chat Orchestration (site da ACL Anthology). A governança não se trata apenas de conformidade; ela também melhora o desempenho do seu sistema de IA agente.

- Otimização — Suas equipes podem criar as mesmas integrações mais de uma vez. Por exemplo, quando cinco equipes diferentes escrevem seu próprio script de consulta de banco de dados para que seu aplicativo de IA se comunique com seus bancos de dados, isso representa cinco vezes o custo de desenvolvimento e cinco conjuntos de bugs a serem mantidos. O MCP permite que você o crie uma vez e o compartilhe com toda a comunidade de engenharia. A economia aumenta à medida que seu número de agentes aumenta.

  Também há um problema de custo por solicitação que a maioria das equipes não percebe no início. Cada definição de ferramenta consome tokens da janela de contexto. Com 20 ferramentas, você está gastando de 5.000 a 10.000 tokens por invocação apenas nas descrições, junto com as consultas do usuário. Isso aumenta a latência e os custos de inferência do LLM e diminui a precisão à medida que o modelo se esforça para escolher a ferramenta certa na lista de ferramentas disponíveis.

  Os agentes que usam pacotes de ferramentas estruturados são aproximadamente três vezes mais precisos nas tarefas do banco de dados do que os agentes que acessam APIs diretamente (para obter mais informações, consulte Middleware for LLMs: Tools Are Instrumental for Language Agents in Complex Environments). A forma como você projeta e apresenta ferramentas para um modelo de IA é importante. Este guia recomenda fornecer às ferramentas esquemas claros, definindo-as para fluxos de trabalho reais em vez de endpoints brutos e limitando as informações na janela de contexto. A seção de estratégia de design de ferramentas MCP deste guia se aprofunda nesses aspectos.
- Segurança e conformidade — imagine um sistema de IA agente que alucina uma etapa de limpeza e tenta excluir um banco de dados de produção. Se o agente herdou as credenciais de

administrador completas do usuário, a exclusão pode ser concluída. Com isolamento de token e credenciais com escopo reduzido que concedem apenas acesso de leitura e criação, ele falha com segurança.

Fluxos de trabalho regulamentados aprimoram ainda mais isso. O guia fornece exemplos (canais de assistência médica que exigem a validação da HIPAA e a anonimização das informações de identificação pessoal antes do processamento dos dados do paciente). A incorporação dessa lógica nas ferramentas do MCP significa que a conformidade sempre acontece de forma determinística.

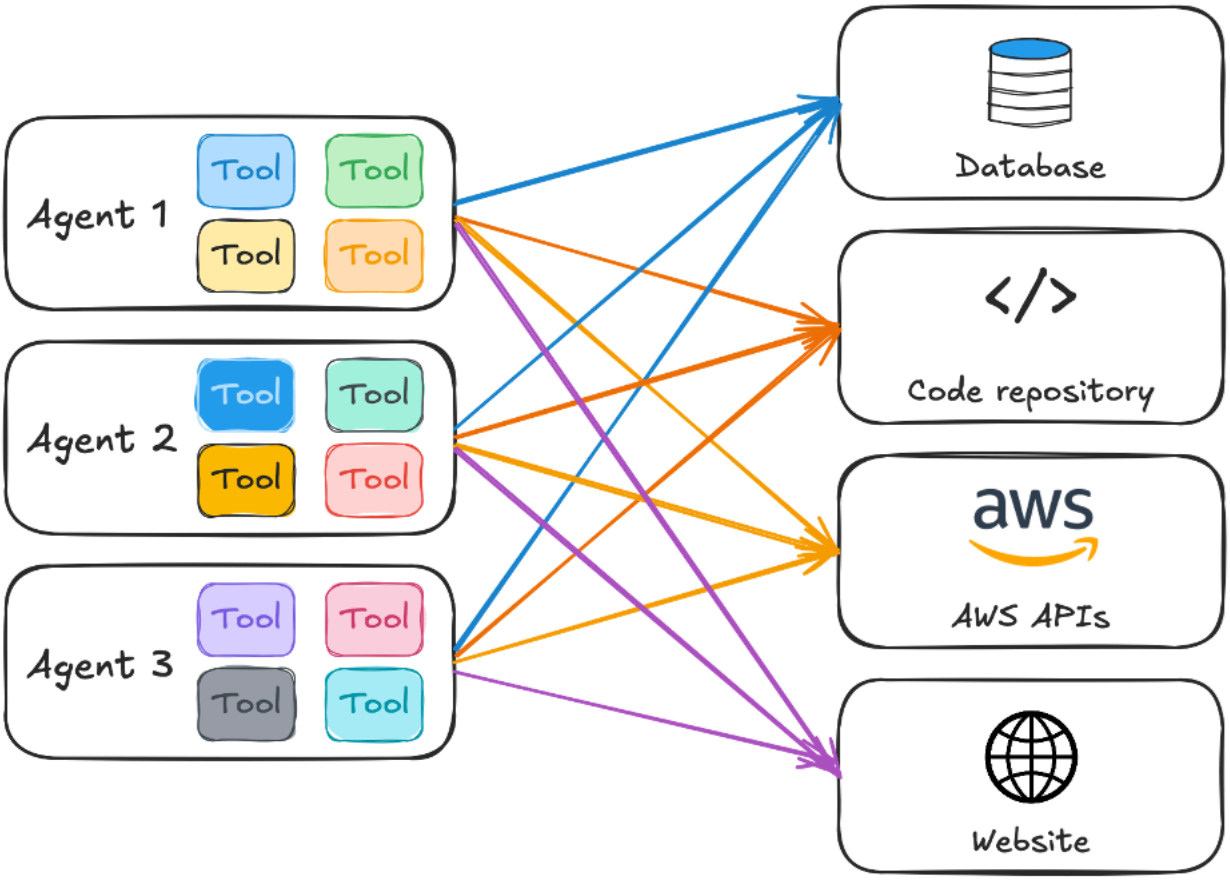
O que é MCP?

LLMs trabalham prevendo uma resposta a uma solicitação com base em seus dados de treinamento. Isso significa que o LLM só pode fornecer respostas sobre dados e eventos que já viu. Métodos como Retrieval Augmented Generation (RAG) e bases de conhecimento permitem que você inclua dados contextuais. No entanto, se você perguntasse a um LLM qual será a previsão do tempo para amanhã ou quantos clientes estão em seu banco de dados, ele provavelmente alucinaria ou não seria capaz de fornecer uma resposta porque isso está fora do conhecimento pré-treinado do LLM. Para poder responder a esses tipos de perguntas, um agente precisa acessar recursos externos, dados e APIs fora do contexto nativo do LLM.

Entendendo as ferramentas

Podemos dar ao LLM acesso a sistemas e contextos adicionais por meio de ferramentas.As ferramentas são funções atribuídas ao LLM para atingir um objetivo claro. Uma ferramenta pode chamar uma API, consultar um banco de dados, realizar operações de calculadora, operar uma caixa de areia de código, realizar uma pesquisa na web e até mesmo invocar outro sistema de IA ou. agent-as-a-tool Cada ferramenta deve incluir uma descrição que diga ao LLM o que a ferramenta faz, quando usá-la e quais parâmetros ela aceita. Isso permite que o LLM tome decisões diferenciadas sobre qual ferramenta ou combinação de ferramentas invocar com base na entrada do usuário.
O LLM é informado sobre quais ferramentas estão disponíveis para o agente, permitindo gerar respostas que instruem o agente a invocar a ferramenta. Por exemplo, quando você pergunta ao LLM quantos clientes estão em seu banco de dados, o LLM envia uma resposta de volta ao agente solicitando a execução da `query_database` ferramenta com parâmetros de entrada específicos.
O LLM determina qual ferramenta invocar e as entradas para a chamada da ferramenta. O agente então executa a ferramenta, que converte a entrada da linguagem natural em uma chamada de função sintaticamente correta e executa a consulta. O agente invoca a ferramenta ou ferramentas com base nas instruções do LLM, e esses resultados são repassados ao LLM. Isso aproveita a capacidade do LLM de raciocinar sobre a entrada baseada em texto e selecionar as ferramentas apropriadas para o trabalho.

A imagem a seguir mostra como cada agente gerencia seu próprio conjunto de ferramentas para cada alvo.



O escalonamento do acesso às ferramentas pode apresentar desafios para soluções de IA agentes:

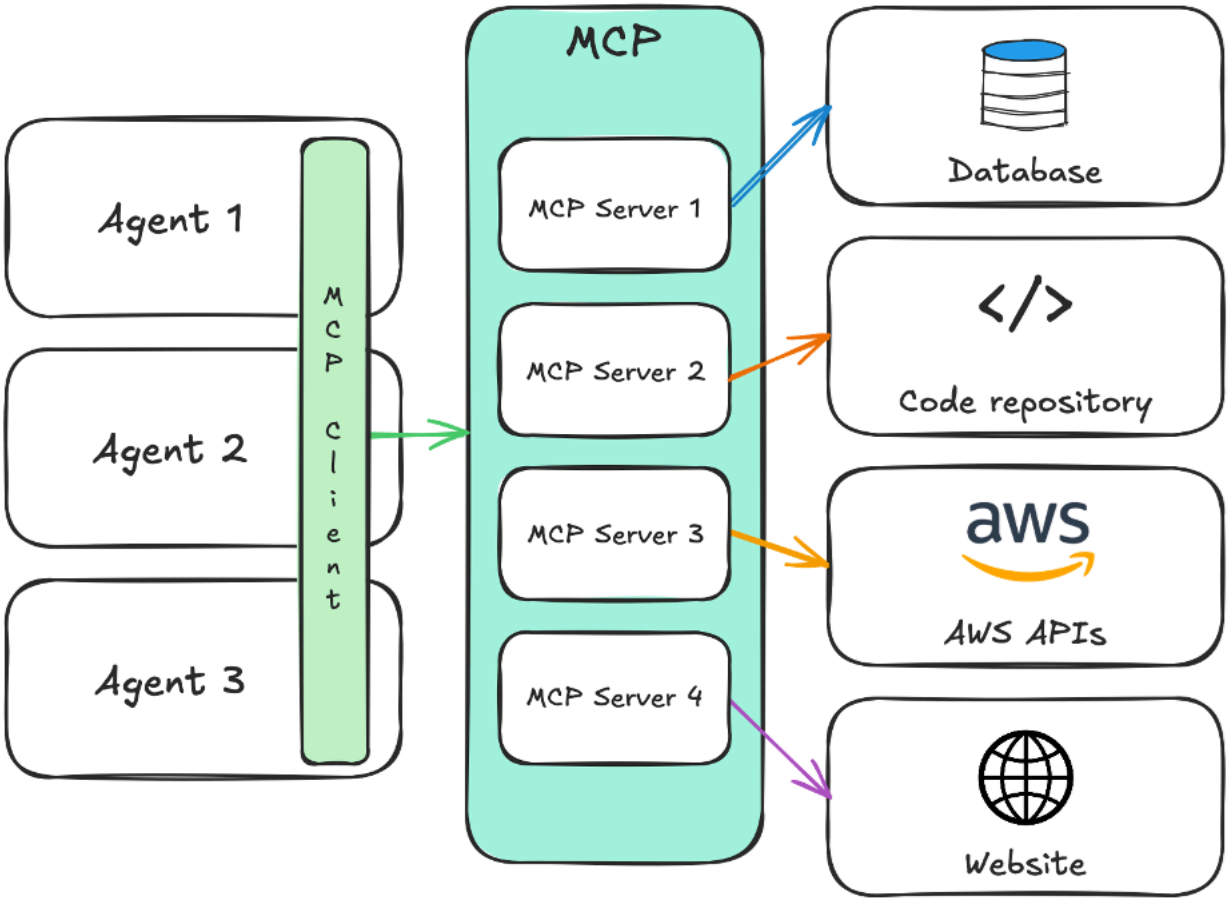
- Se cada desenvolvedor está criando sua própria ferramenta para os mesmos recursos externos, há muito esforço duplicado e formas não padronizadas de interagir com esses recursos externos. Isso produz implementações inconsistentes em seus agentes. Embora você possa resolver esse problema desenvolvendo ferramentas padrão em bibliotecas e distribuindo-as, isso carece de governança centralizada. Isso dificulta a aplicação de políticas de segurança, o rastreamento do uso de ferramentas, o gerenciamento de versões entre as equipes ou a garantia da conformidade com os padrões organizacionais. Além disso, ao incorporar ferramentas diretamente ao agente, você deve reimplantar seu agente sempre que uma nova ferramenta for criada ou uma existente for atualizada.
- Fornecer ferramentas para um LLM consome sua janela de contexto. A janela de contexto é o número de tokens (unidades de texto que são LLMs processadas — normalmente representando palavras, partes de palavras ou pontuação) que um modelo pode considerar a qualquer momento. LLMs têm limites de janela de contexto. As ferramentas e sua documentação consomem essa janela de contexto finita junto com as solicitações do sistema e as solicitações do usuário. À medida que a janela de contexto é preenchida, o desempenho LLMs pode ser reduzido devido

a vários fatores: dificuldade em identificar informações relevantes, aumento da complexidade do processamento e redução da capacidade de raciocínio. O desafio é agravado quando as definições de ferramentas, as solicitações do sistema e o histórico de conversas competem por um espaço limitado na janela de contexto porque são fornecidos em cada invocação do LLM.

Assim, o número de ferramentas e a forma como elas são documentadas têm impacto direto no desempenho do LLM, como tempo de resposta e precisão.

O MCP estabelece um padrão universal para conectar agentes a recursos externos. É comumente chamado de “USB-C para aplicativos de IA”. Em vez de registrar ferramentas diretamente com agentes, os servidores MCP atuam como intermediários para hospedar ferramentas que são descobertas e invocadas por meio do JSON-RPC 2.0. Em vez de adicionar dezenas ou centenas de ferramentas diferentes ao seu agente e mantê-las ao longo do tempo, o MCP permite que você registre servidores MCP que encapsulam as ferramentas que seu agente pode acessar. Essa abordagem padroniza como as ferramentas são empacotadas, apresentadas e invocadas. Isso pode ajudar a enfrentar os desafios de escala e governança do uso de ferramentas em seus agentes. Ele também separa o desenvolvimento e as operações de agentes das ferramentas que usa para recursos externos.

A figura a seguir mostra agentes usando MCP para acessar recursos externos.



No entanto, o padrão MCP não resolve todos os desafios de escalabilidade e governança. A implementação de servidores MCP deve ser combinada com estratégias eficazes de design de ferramentas, hospedagem e governança corporativa. Este guia fornece as melhores práticas para cada estratégia para ajudá-lo a criar e usar o MCP como parte de suas soluções de IA para agentes.

Quando usar o MCP

O MCP fornece infraestrutura estratégica para escalar suas iniciativas de inteligência artificial. Ao centralizar o gerenciamento e a governança de ferramentas, os servidores MCP reduzem o custo cumulativo de criar e manter integrações personalizadas entre vários agentes. Isso proporciona retornos crescentes à medida que seu ecossistema de agentes se expande.

O MCP provavelmente se torna parte de sua estratégia quando:

- Você precisa de uma governança centralizada sobre como os agentes acessam sistemas e serviços corporativos, como bancos de dados APIs, ferramentas internas e integrações de terceiros.

- Os desenvolvedores passam muito tempo escrevendo integrações personalizadas que não são consistentes em todas as implementações.
- Você tem ferramentas duplicadas que podem servir a recursos comuns.
- Você deseja oferecer suas ferramentas ou dados proprietários a consumidores externos ou sistemas de agentes terceirizados por meio de interfaces MCP padronizadas e governadas, liberando novos fluxos de receita e mantendo a segurança e o controle.

Depois de decidir que os servidores MCP farão parte de sua estratégia, avalie se as implementações de servidores MCP de código aberto existentes atendem às suas necessidades, se elas precisam ser aprimoradas ou se você precisa criar servidores personalizados. Muitas implementações de servidor MCP pré-criadas estão disponíveis em repositórios públicos e abrangem recursos comuns, como acesso ao sistema de arquivos, navegação na web, sandboxes de código, acesso ao banco de dados e integrações de API.

Em muitos casos, servidores MCP preexistentes são suficientes. Por exemplo, AWS fornece o Servidor AWS MCP, um servidor MCP remoto gerenciado que fornece aos assistentes e agentes de IA acesso seguro e autenticado Serviços da AWS por meio de interações em linguagem natural. Você pode usar o Servidor AWS MCP para realizar AWS tarefas complexas de várias etapas combinando acesso em tempo real à AWS documentação, chamadas de API sintaticamente corretas e fluxos de trabalho pré-criados chamados Agente SOPs que seguem as melhores práticas. AWS AWS testa continuamente os Servidor AWS MCP para garantir que os agentes do cliente possam usá-los com sucesso.

Você deve testar esses servidores MCP existentes com seus agentes para determinar se eles atendem aos seus casos de uso. Se um agente não conseguir concluir fluxos de trabalho, gerar respostas incorretas ou inadequadas, não conseguir navegar por processos complexos de várias etapas ou perder importantes práticas recomendadas ou considerações de segurança específicas do domínio, você deve considerar aprimoramentos em várias dimensões.

Quando os servidores MCP existentes não atendem totalmente às suas necessidades e têm dificuldade em usar as ferramentas existentes corretamente ou produzir respostas precisas, considere estas abordagens de aprimoramento antes de criar servidores personalizados:

- Enriqueça o contexto do agente — Se seu agente tiver dificuldade em usar as ferramentas de forma correta ou eficiente em um servidor MCP existente, considere complementar essas definições de ferramentas com documentação ou exemplos adicionais. Isso ajuda a fornecer contexto adicional ao LLM.

- Adicione ferramentas complementares — amplie os servidores MCP existentes com ferramentas que acessam dados ou contextos organizacionais adicionais de que os agentes precisam para concluir fluxos de trabalho com êxito.
- Melhore o subjacente APIs — simplifique seu serviço APIs para torná-lo mais compatível com o LLM, reduzindo a complexidade dos parâmetros, fornecendo mensagens de erro mais claras e oferecendo padrões sensatos que os agentes possam usar.

Embora o uso de implementações de servidores MCP existentes acelere o desenvolvimento de recursos comuns, a criação de servidores MCP personalizados é uma necessidade quando seu caso de uso exige funcionalidade especializada. Os servidores MCP personalizados ajudam você a encapsular a experiência do domínio, aplicar padrões organizacionais, melhorar a confiabilidade dos agentes para fluxos de trabalho complexos e oferecer suporte à conformidade com os requisitos de segurança. Considere criar um servidor MCP personalizado nas seguintes situações:

- Fluxos de trabalho específicos do domínio — Os fluxos de trabalho de várias etapas que exigem experiência no domínio devem ser encapsulados em ferramentas MCP personalizadas quando o conhecimento necessário não é capturado na documentação da API. Por exemplo, em vez de permitir que os agentes organizem canais complexos de dados de saúde que devem validar a conformidade com a Lei de Portabilidade e Responsabilidade de Seguros de Saúde (HIPAA), anonimizar as PII e transformar para o formato HL7 FHIR, forneça uma ferramenta que incorpore diretamente a experiência do domínio. `process_patient_data` Isso elimina a dependência do LLM para orquestrar e executar corretamente as etapas do fluxo de trabalho, o que melhora a consistência e a conformidade.
- Abstrações do caminho dourado — Os agentes podem ter dificuldade em implementar abordagens ideais porque não têm contexto organizacional e adotam padrões básicos em vez das melhores práticas organizacionais. Nesses cenários, você pode aplicar padrões prescritivos de custo, desempenho ou segurança encapsulando esses caminhos dourados em ferramentas MCP personalizadas. Por exemplo, em vez de permitir que os agentes implantem uma infraestrutura com configurações padrão que podem ser inseguras ou ineficientes, forneça uma `deploy_secure_infrastructure` ferramenta que incorpore diretamente os padrões da sua organização.
- Orquestração complexa de vários serviços — Em vez de fazer com que o agente orquestre fluxos de trabalho complexos tentando inferir a sequência correta e o conjunto de serviços a serem usados em cada etapa, você pode criar de forma determinística essa lógica dentro de uma ferramenta MCP. Talvez você também queira fornecer experiência sobre os padrões ideais de

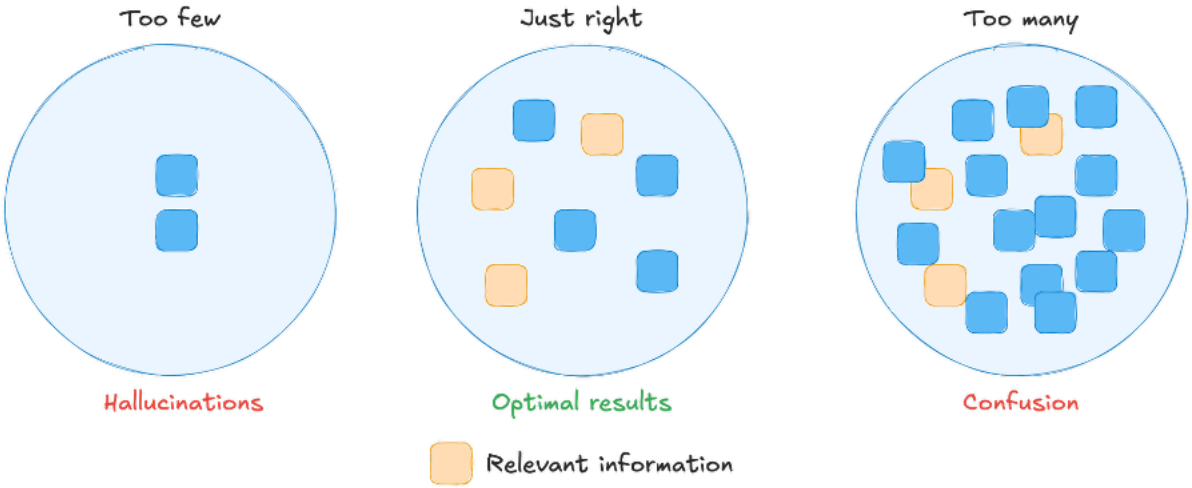
integração de serviços que o agente talvez não conheça. Isso também pode melhorar a precisão e a eficiência de seus agentes.

- Práticas recomendadas específicas do serviço — Isso é comum para ferramentas focadas na segurança que ajudam os agentes a implementar políticas de criptografia, controles de acesso e padrões de conformidade específicos para o serviço que está sendo acessado por meio da ferramenta do agente. Além disso, se houver melhores práticas operacionais específicas do serviço que não sejam óbvias, o uso de um servidor MCP pode ajudá-lo a garantir que elas sejam implementadas e não deixadas para um agente raciocinar.

Estratégia de design de ferramentas MCP

A principal tarefa do cliente e servidor MCP é descobrir e apresentar ferramentas ao LLM para que ele possa usá-las para melhorar suas respostas. Isso faz do design da ferramenta MCP uma das estratégias mais importantes para criar soluções MCP eficazes. Do ponto de vista do modelo, as ferramentas são uma função que eles podem invocar conforme necessário para fornecer respostas mais precisas e completas. A interface da função abstrai a implementação subjacente de uma ferramenta, que pode variar de um invólucro em torno de uma única chamada de API até uma lógica complexa de fluxo de trabalho.

No entanto, você deve encontrar um equilíbrio com a quantidade de ferramentas fornecidas ao LLM. Se houver poucas ferramentas, o LLM pode não conseguir coletar o contexto e as informações corretos, portanto, fará a melhor suposição com as informações disponíveis no modelo. Se houver muitas ferramentas, o LLM pode ficar confuso sobre a seleção e a sequência corretas das ferramentas, levando a alucinações. Seu objetivo é obter o número certo de ferramentas. A imagem a seguir mostra os desafios de poucas e muitas ferramentas.



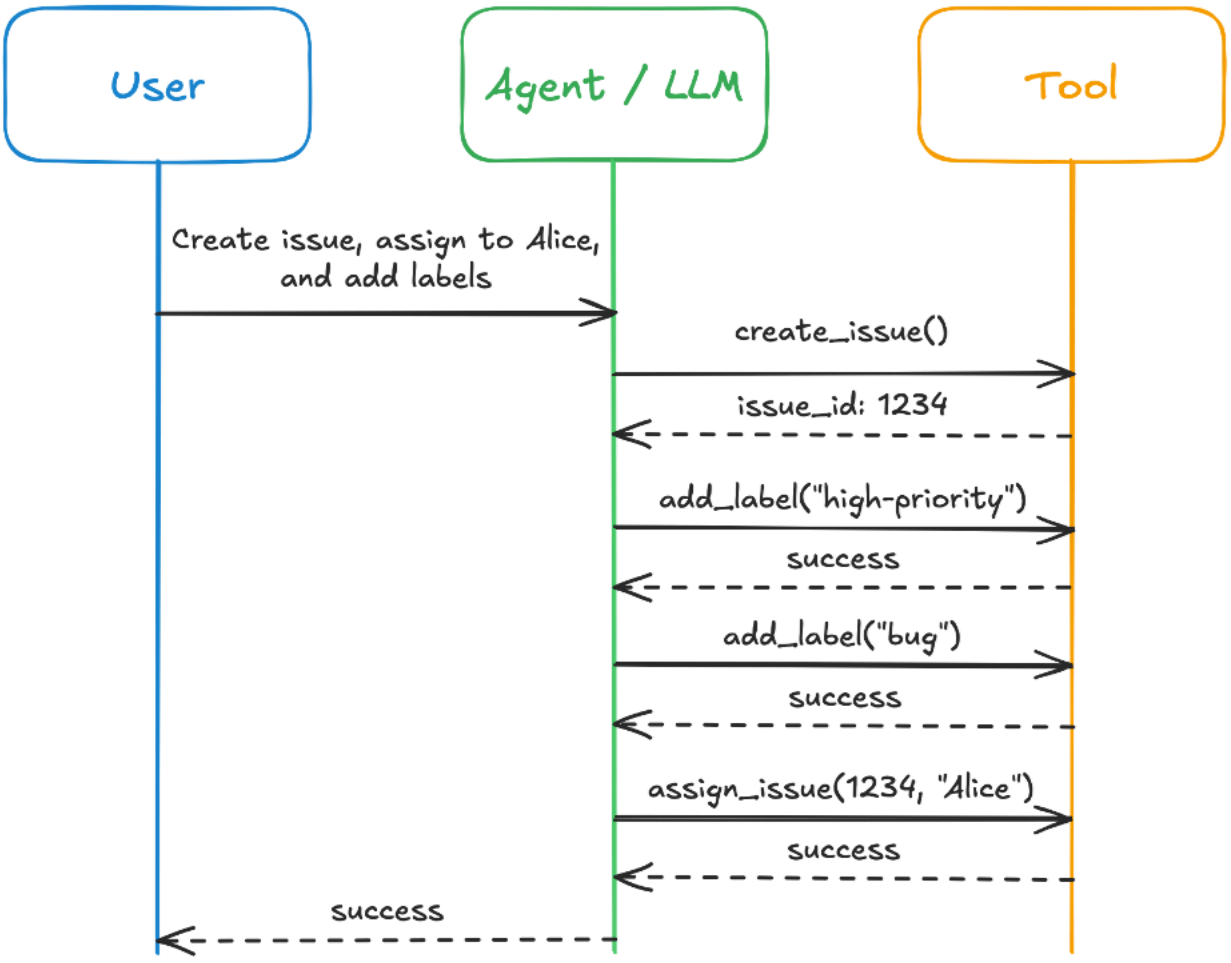
A solução requer a compreensão de quantas ferramentas fornecer e como definir o escopo de cada ferramenta. A granularidade de suas ferramentas, sejam elas mapeadas para chamadas de API individuais ou fluxos de trabalho completos, afeta diretamente o número total de ferramentas que os agentes precisam e a eficiência com que podem usá-las. Esta seção fornece as melhores práticas para definir o escopo das ferramentas MCP, criar definições de ferramentas, descobrir ferramentas e organizá-las.

Escopo da ferramenta

Há duas abordagens para o desenvolvimento de ferramentas: granular e granulada.

Granular

Em uma abordagem granular, você criaria uma ferramenta por API, ação ou consulta. Por exemplo, você pode criar`create_issue`,`get_issue`, `add_labelassign_issue`, e `close_issue` ferramentas para seu repositório Git. Isso permitiria que o LLM fizesse chamadas granulares para cada API e orquestrasse cada uma conforme necessário. Considere o seguinte aviso: “Crie um problema para o serviço do produto chamado 'A consulta retorna apenas resultados parciais', rotule-o como um bug e de alta prioridade e atribua-o a Alice”. A imagem a seguir mostra como uma tool-per-API abordagem responderia a essa solicitação.

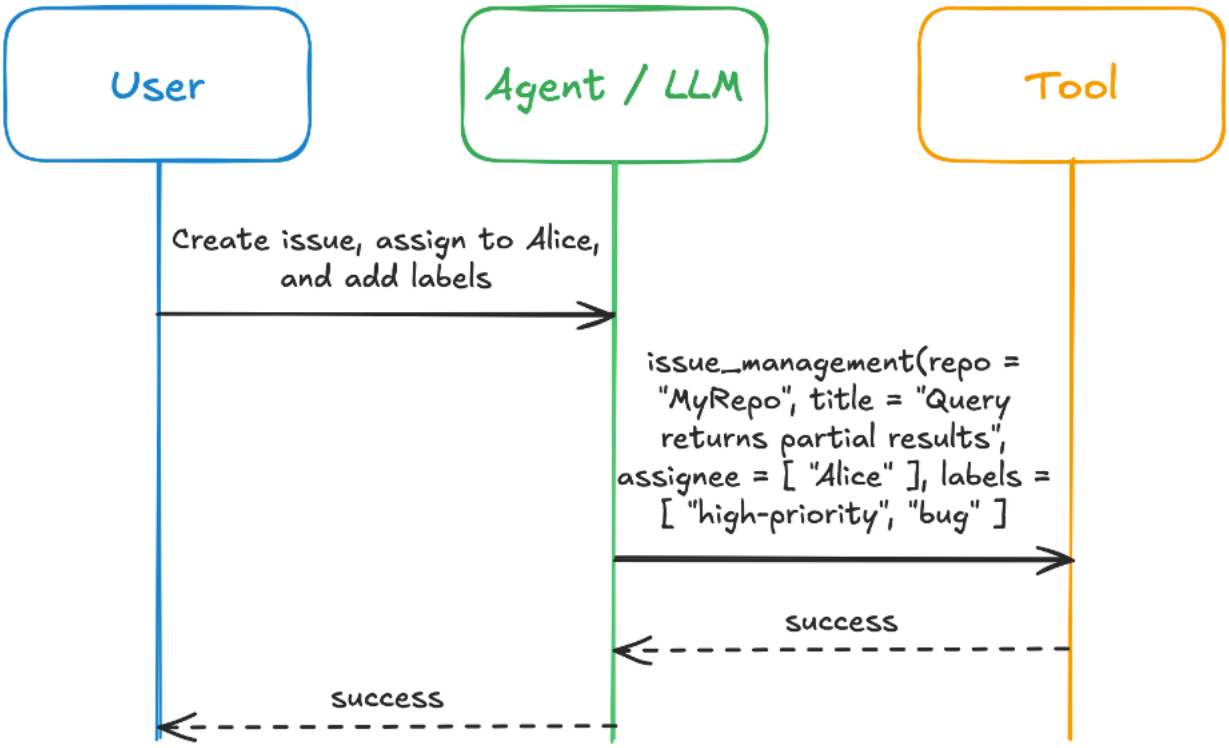


Nessa abordagem, o prompt do sistema e cada definição de ferramenta registrada são fornecidos ao LLM em cada chamada. Isso consome contexto adicional e incorre em uma penalidade de latência

porque cada chamada de ferramenta representa uma chamada individual para o LLM. Também aumenta a complexidade de lidar com erros no fluxo de trabalho.

Baixa granularidade

Uma abordagem granulada ou orientada por fluxo de trabalho seriam ferramentas orientadas ao fluxo de trabalho. A ferramenta se concentra na intenção end-to-end do usuário em vez da estrutura da API. Em vez de uma tool-per-API, você tem uma ferramenta que chama muitos de forma determinística. APIs Usando o exemplo anterior do repositório Git, você pode criar uma `create_and_setup_issue` ferramenta que é chamada uma vez pelo agente. A implementação da ferramenta cria o problema, adiciona rótulos e o atribui a um usuário, com base nos parâmetros fornecidos à ferramenta. A imagem a seguir mostra como uma abordagem granulada processaria o mesmo prompt.



Essa abordagem mostra como toda a complexidade permanece oculta da camada LLM. Quando a lógica de orquestração é incorporada à implementação da ferramenta, todas as etapas sequenciais, o registro, a lógica de repetição, os disjuntores e a limitação de taxa são executadas de forma determinística na ferramenta. A abordagem orientada pelo fluxo de trabalho torna mais simples para o LLM invocar a ferramenta correta com os parâmetros certos. É importante observar que algumas APIs podem já fornecer a intenção do fluxo de trabalho, como a API do Amazon `RunInstances`

EC2. Nesses casos, a tool-per-API pode fornecer o design orientado ao fluxo de trabalho que você deseja.

No entanto, as ferramentas também podem se tornar muito grossas. Se sua única ferramenta de fluxo de trabalho tentar fazer muitas coisas e tiver muitos parâmetros possíveis, o LLM poderá ter dificuldade em raciocinar sobre como usar a ferramenta corretamente. Também pode criar desafios com a seleção de parâmetros e o tratamento de erros. Portanto, o desenvolvimento de ferramentas deve encontrar um equilíbrio que se alinhe à intenção do usuário e evite pouca ou muita funcionalidade em uma única ferramenta. Recomendamos que você crie ferramentas com base em fluxos de trabalho completos do usuário, agrupando operações que normalmente ocorrem juntas (como três ou mais chamadas de API). Também recomendamos que você decomponha ferramentas que excedam oito ou mais parâmetros ou lidem com várias intenções de usuário distintas. Teste com instruções reais para verificar se os agentes podem usar cada ferramenta corretamente.

Se você tiver fluxos de trabalho complexos e dinâmicos que não podem ser facilmente encapsulados como uma ferramenta determinística, considere usar o padrão. agent-as-tool Em vez de seu agente principal tentar orquestrar tarefas complexas em um fluxo de trabalho, um agente especializado pode atuar como uma ferramenta. Esses tipos de ferramentas podem implementar tomadas de decisão e ramificações avançadas, além de lidar com erros e repetir a lógica que não pode ser facilmente gerenciada em código determinístico. Isso é semelhante, mas distinto do protocolo Agent2Agent (A2A). O protocolo A2A é complementar, fornecendo interoperabilidade e colaboração entre agentes em qualquer estrutura de agente.

Recomendamos que você comece com a análise do fluxo de trabalho mapeando os fluxos de trabalho mais comuns dos usuários para identificar os principais recursos de que cada agente precisa. Isso estabelece seu conjunto mínimo de ferramentas viável. Com base em nossa experiência no desenvolvimento de servidores MCP em grande escala, recomendamos as seguintes práticas. Quando essas práticas entrarem em conflito, priorize a intenção do usuário e o fluxo de trabalho.

Melhores práticas para o escopo da ferramenta MCP

- Pense em histórias de usuários e agrupe operações comuns — as ferramentas devem ser mapeadas diretamente para concluir as interações do usuário, em vez de exigir a orquestração de várias operações. Se os fluxos de trabalho normalmente exigirem três ou mais chamadas separadas, combine-as em uma única ferramenta. Isso reduz a carga cognitiva no LLM, minimiza o número de chamadas de ferramentas, reduz o consumo de contexto e a latência necessários para concluir tarefas e melhora a precisão e a latência.

- Limite os parâmetros a oito ou menos — Se uma ferramenta exceder oito parâmetros, decomponha-a em várias ferramentas. LLMs lutam com a seleção de parâmetros à medida que a complexidade aumenta.

> **Note**
>
> Se as operações de agrupamento exigirem mais de oito parâmetros, priorize o agrupamento em vez da contagem de parâmetros, pois simplificar o fluxo de trabalho é mais valioso do que limites estritos de parâmetros.

- Operações separadas de leitura e gravação — forneça ferramentas diferentes para ler dados e modificá-los. Essa separação torna explícito quando os agentes estão realizando operações potencialmente destrutivas, permite políticas de autorização diferentes e reduz o risco de modificações não intencionais durante a coleta de informações.
- Forneça padrões razoáveis — Crie ferramentas para que o LLM precise especificar somente os parâmetros específicos da solicitação individual. Os padrões reduzem a complexidade dos parâmetros e melhoram a precisão da seleção de ferramentas, minimizando as informações sobre as quais o LLM deve raciocinar.
- Prefira a execução determinística — Torne a execução da ferramenta e a saída determinísticas quando possível. As ferramentas determinísticas são mais confiáveis e fáceis de testar. Para fluxos de trabalho complexos que exigem orquestração inteligente, lógica de ramificação ou tratamento avançado de erros que não podem ser facilmente gerenciados em código determinístico, considere usar agentes especializados como ferramentas. No entanto, use esse padrão seletivamente porque ele adiciona complexidade.

Definições de ferramentas

Quando um LLM recebe uma solicitação que não pode tratar diretamente, ele revisa as ferramentas disponíveis para ajudá-lo a concluir a solicitação. O LLM seleciona ferramentas com base em sua compreensão semântica dos nomes e descrições das ferramentas fornecidas e em todas as instruções fornecidas no prompt. Em seguida, ele criará uma entrada com base no esquema de entrada definido e esperará uma saída com base no esquema de saída. Portanto, criar definições descritivas de ferramentas e esquemas de entrada e saída validados é fundamental para ajudar o LLM a selecionar ferramentas de forma eficaz. Geralmente, há duas abordagens para criar essa documentação: a abordagem de especificação da ferramenta e a abordagem docstring.

Abordagem de especificação de

A abordagem recomendada é seguir diretamente a especificação da ferramenta MCP ao definir a ferramenta. O exemplo a seguir é mostrado usando o decorador de ferramentas Strands Agent:

```
@tool(
  name = "search_website",
  description = "This tool searches the provided website for semantic matches to the
 query provided",
  inputSchema = {
    "json": {
      "type": "object",
      "properties": {
        "url": {
          "type": "string",
          "description": "The url of the website to load and search."
        },
        "query": {
          "type": "string",
          "description": "The content you want to try and match in the website."
      }
    },
    "required": ["url", "query"]
  },
  outputSchema = {
    "json": {
      "type": "object",
      "properties": {
      "results": {
          "type": "array",
          "items": {
            "type": "string"
          }
        }
      }
    }
  }
)
def search_website:
  …
```

O uso de campos padrão, comoname,`description`,`inputSchema`, e `outputSchema` garante que cada ferramenta tenha uma documentação consistente que tanto o LLM quanto os humanos possam

entender. Cada ferramenta deve definir esses campos no mínimo e, opcionalmente, fornecer um título e anotações, que são dicas opcionais sobre o comportamento da ferramenta. Quando possível, use enums para valores de parâmetros para facilitar a seleção das opções corretas pelo LLM. As enumerações funcionam melhor para conjuntos finitos, como valores de status ou prioridade, mas não são adequadas para texto de formato livre, valores dinâmicos, números arbitrários ou identificadores de recursos. Nesses casos, forneça descrições e exemplos claros. Inclua também um valor padrão quando possível para que o LLM não precise adivinhar qual é a opção correta. Lembre-se de que as definições da ferramenta estão incluídas no prompt do LLM em cada invocação, consumindo espaço na janela de contexto junto com as instruções do sistema e o histórico de conversas.

Abordagem Docstring

Outra abordagem, se você estiver escrevendo suas ferramentas em Python, é usar docstrings para fornecer a descrição, o uso e a saída da ferramenta. Veja a seguir um exemplo dessa abordagem:

```
def search_website(url: str, query: str) -> list:

    """
    This tool loads the specified website and then attempts to find content that
 matches the provided query through semantic search. It provides back a list of strings
 that are the sentences that match the query.
    Args:
      url: the website url to load
      query: the content you want to semantically match in the website
    """
```

Docstrings não impõem um esquema ou formato padronizado. O uso dessa abordagem pode gerar resultados inconsistentes com base em como os desenvolvedores de ferramentas escolhem documentar cada ferramenta. Definir e aplicar um padrão para toda a organização é essencial se você seguir essa abordagem.

Melhores práticas para definições de ferramentas MCP

- Siga as especificações da ferramenta MCP — Forneça `name`,`description`,`inputSchema`, e `outputSchema` campos para cada ferramenta. Para implementações de Python, use modelos Pydantic para fornecer documentação embutida por meio de descrições de campo, validação automática de tipos e valores restritos por meio de enumerações. Isso torna os esquemas autodocumentados e melhora a compreensão do LLM sobre as opções de parâmetros válidos.

- Escreva descrições como instruções — As descrições das ferramentas são instruções que orientam a tomada de decisão do LLM. Inclua os componentes essenciais da finalidade da ferramenta (o que a ferramenta faz), quando usá-la (cenários ou padrões de intenção do usuário), o contexto da saída (para que a saída é usada), os parâmetros e as condições de erro.
- Forneça exemplos concretos — Incluir exemplos de fluxo de trabalho com valores reais é a maneira mais eficaz de orientar LLMs sobre o uso correto da ferramenta.
- Documente as dependências de forma explícita — inclua pré-requisitos, sequências numeradas, mudanças de estado e ações de acompanhamento.

Descoberta de ferramentas

Há três abordagens para descobrir e registrar ferramentas em seu agente com servidores MCP: definição estática, descoberta dinâmica e função de pesquisa.

Definição estática

Primeiro, você pode definir estaticamente as ferramentas disponíveis diretamente no código do agente. Nessa abordagem, você define uma ferramenta remota (um objeto de referência do lado do cliente em uma estrutura como o Strands Agent SDK) para cada ferramenta fornecida pelo servidor MCP que é acessada por um cliente MCP. O exemplo a seguir usa transporte HTTP simplificável:

```
from mcp.client.streamable_http import streamablehttp_client
from strands import Agent
from strands.tools.mcp import MCPClient

streamable_http_mcp_client = MCPClient(
    lambda: streamablehttp_client("https://mcp1:8000/mcp")
)

reverse_text = RemoteTool(
    name="reverseText",
    client=streamable_http_mcp_client
)

agent = Agent(tools=[reverse_text])
```

O registro individual de ferramentas ajuda você a ser muito seletivo quanto às ferramentas que você disponibiliza para o LLM, o que minimiza a quantidade de janela de contexto usada. A desvantagem

é que isso requer o conhecimento dos nomes das ferramentas disponíveis e pode ser frágil se as ferramentas disponíveis mudarem no servidor MCP.

Descoberta dinâmica

A próxima abordagem é usar a descoberta dinâmica e registrar todas as ferramentas disponíveis com o agente. Essa abordagem consome o contexto linearmente à medida que mais ferramentas são adicionadas ao servidor MCP. Veja a seguir um exemplo dessa abordagem:

```
from mcp.client.streamable_http import streamablehttp_client
from strands import Agent
from strands.tools.mcp import MCPClient

streamable_http_mcp_client = MCPClient(
    lambda: streamablehttp_client("https://mcp1:8000/mcp")
)

with streamable_http_mcp_client:
    tools = streamable_http_mcp_client.list_tools_sync()
    agent = Agent(tools=tools)
```

Considere um cenário em que uma definição de ferramenta típica consome aproximadamente 250 a 500 tokens (incluindo nome, descrição e esquema). O registro de 20 ferramentas consumiria de 5.000 a 10.000 tokens da sua janela de contexto. Quando você tem um pequeno número de servidores MCP e tem controle sobre o número de ferramentas, essa opção é a mais simples de implementar. No entanto, se se espera que a lista de ferramentas cresça, isso pode criar problemas silenciosos de gerenciamento de contexto em seus agentes. Uma variação alternativa dessa abordagem é usar um parâmetro de filtro de ferramentas ao chamar`list_tools`, como o que o SDK do Strands Agents fornece, para reduzir o número de ferramentas registradas com o agente.

Função de pesquisa

A terceira opção é usar uma função de pesquisa para encontrar ferramentas relevantes durante o tempo de execução. Você lista todas as ferramentas disponíveis do seu servidor MCP e, em seguida, executa uma pesquisa semântica sobre essas ferramentas com base no prompt do usuário. Em seguida, as ferramentas resultantes são registradas com seu agente. O Amazon Bedrock AgentCore Gateway fornece um recurso de pesquisa semântica nativa que pode facilitar a implementação desse tipo de solução.

Práticas recomendadas para descoberta de ferramentas MCP

- Preservação da janela de contexto — Escolha uma abordagem de descoberta e registro de ferramentas que conserve o máximo possível da janela de contexto.
- Use recursos de filtragem de ferramentas ou pesquisa semântica — Forneça dinamicamente ao LLM um conjunto de ferramentas com escopo reduzido para escolher, o que melhora sua precisão e eficácia na escolha da ferramenta certa. A filtragem de ferramentas pode operar em nomes de ferramentas (correspondência ou padrões exatos), descrições de ferramentas (correspondência semântica) ou tags de domínio ou categoria. A pesquisa semântica é particularmente eficaz para comparar a intenção do usuário com as descrições das ferramentas. Ambas as abordagens reduzem o uso da janela de contexto.

Organização de ferramentas

Descobrir as ferramentas certas e garantir que o LLM possa usá-las de forma eficaz é uma das partes mais críticas do desenvolvimento eficaz de ferramentas. Ao começar a desenvolver servidores MCP, você precisa de uma estratégia que determine:

- Quantas ferramentas entram em um servidor MCP
- Quais ferramentas não devem ser colocadas no mesmo servidor MCP
- Como nomear ferramentas para torná-las pesquisáveis e evitar colisões de nomes (ferramentas diferentes com o mesmo nome)
- Como documentar as ferramentas e o servidor MCP para torná-los fáceis de usar pelo LLM

A organização do namespace é um padrão de design que evita colisões de nomes de ferramentas, agrupa funcionalidades relacionadas e facilita a identificação eficiente de ferramentas por. LLMs O padrão estabelece uma categorização estruturada que é análoga aos sistemas de armazenamento organizados, em vez da acumulação não estruturada. Recomendamos o domain-noun-verbpadrão para nomeação de ferramentas. Por exemplo, `github_issue_creategithub_issue_list`,`github_issue_update`,`github_pullrequest_creat` ou`github_pullrequest_merge`. A vantagem desse padrão é evidente ao examinar o comportamento de classificação alfabética. Quando as ferramentas são listadas em ordem alfabética, todas as operações relacionadas a problemas se agrupam (`create`,,`update`)`list`, seguidas pelas operações de pull request (,,). `create` `list` `merge` O substantivo (tipo de recurso) funciona como um limite organizacional. Essa estrutura facilita tanto a digitalização da ferramenta

LLM quanto a navegação na documentação humana porque as funcionalidades relacionadas se agrupam naturalmente.

O servidor MCP deve ser limitado no nível do domínio, mas pode ser subdividido com base na separação de tarefas dos recursos que ele fornece. Por exemplo, você pode ter servidores MCP separados para operações de gravação e leitura em um banco de dados. Para impor essa separação, é recomendável implementar grades de proteção no nível do agente que restrinjam quais servidores MCP podem ser acessados com base na intenção e nas permissões do usuário. Isso pode ser obtido por meio de uma combinação das seguintes opções:

- Carregamento condicional do servidor — Carregue o servidor MCP somente para leitura quando o agente detectar operações de leitura na entrada do usuário.
- Filtragem baseada em permissões — Use a autorização do usuário para conceder acesso somente aos servidores MCP apropriados.

Finalmente, você desejará criar um limite superior no número de ferramentas fornecidas por um servidor MCP. Não faça suposições sobre como os agentes usarão seu servidor MCP. Eles podem listar ingenuamente todas as ferramentas disponíveis e fornecê-las todas ao LLM. Se você tiver mais de 50 ferramentas em um único servidor, considere dividi-las em vários servidores.

Melhores práticas para organização de ferramentas de MPC

- Use o padrão domain-noun-verb de nomenclatura para ferramentas — Implemente estratégias para evitar colisões de nomes nos servidores MCP e nos agentes.
- Defina um limite superior — Restrinja o número de ferramentas em um único servidor MCP.
- Divida os servidores MCP — Use a separação de tarefas para dividir os servidores MCP em grupos lógicos.

Estratégia de hospedagem MCP

A abstração das ferramentas disponíveis em servidores MCP separa o desenvolvimento de seu agente das ferramentas disponíveis. Isso apresenta os desafios de onde você hospeda seu servidor MCP e como as ferramentas são organizadas dentro desses servidores.

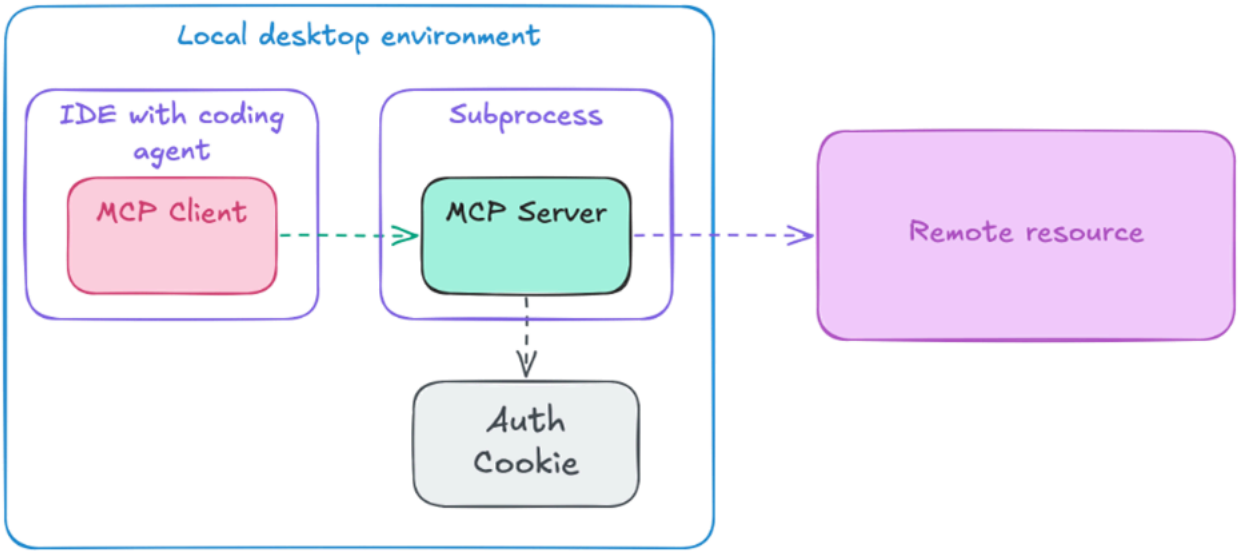
Abordagens de hospedagem

Há três opções para hospedar seus servidores MCP: executá-los localmente em uma máquina de usuário final, hospedá-los remotamente ou hospedá-los por meio de um gateway MCP. Cada opção tem vantagens e desvantagens.

Hospedagem local

A hospedagem local executa o servidor MCP como um subprocesso em sua máquina local junto com o agente que se comunica com o servidor usando JSON-RPC em fluxos de entrada e saída padrão. Essa abordagem não exige autenticação entre o cliente e o servidor. As ferramentas podem interagir com aplicativos e arquivos locais, usar credenciais armazenadas localmente e herdar o acesso à rede da máquina local do usuário. Esse é o padrão de hospedagem mais simples e tem vários benefícios.

Muitos clientes começam a usar o MCP usando servidores locais. Eles permitem que os engenheiros iterem e resolvam rapidamente uma variedade de problemas em seu ambiente local. Considere um servidor MCP que se conecta a um repositório Git que o assistente de codificação de um engenheiro está usando. Manter o servidor MCP local faz muito sentido porque ele pode usar as credenciais exclusivas do engenheiro para acessar o repositório e não adiciona uma chamada de rede extra a um servidor MCP remoto. A imagem a seguir mostra um servidor MCP hospedado localmente sendo usado com um agente de codificação em um IDE.



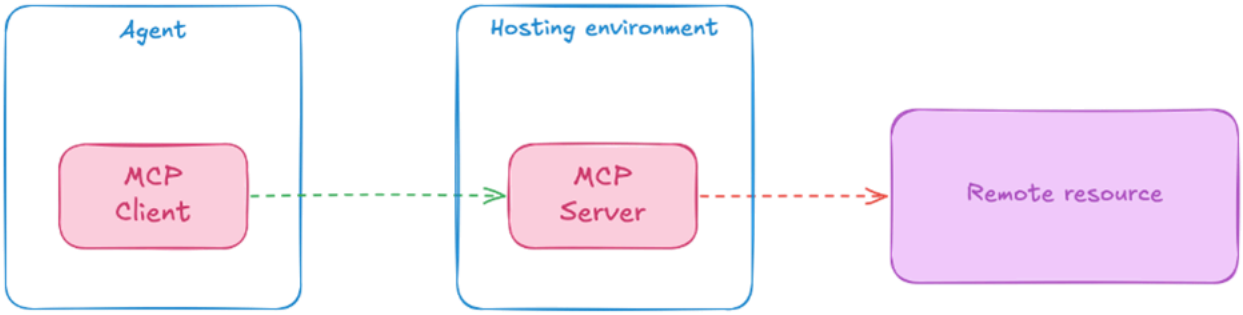
Para esses tipos de implantações, você deve considerar como os servidores MCP são desenvolvidos e distribuídos. A maioria dos clientes desenvolve um registro MCP em que os servidores podem ser registrados e baixados pelos usuários finais. É muito semelhante a um registro de contêiner em que um usuário pode pesquisar recursos específicos e encontrar os servidores MCP adequados às suas necessidades.

Existem registros públicos de MCP, como o Registro Oficial de MCP, e há registros hospedados de forma privada. Normalmente, as organizações alinham sua estratégia de registro de MCP às políticas existentes sobre distribuição de software de código aberto, registros de contêineres e gerenciamento interno de pacotes. Você deve considerar fatores como verificação de segurança, fluxos de trabalho de aprovação e requisitos de conformidade.

No entanto, a hospedagem local apresenta desafios operacionais que as organizações devem considerar. Primeiro, os usuários finais devem descobrir, baixar e configurar os servidores MCP de forma independente. Isso pode aumentar a complexidade para começar a usar cada servidor MCP individual que eles usam localmente. Segundo, você não pode controlar o ciclo de vida do servidor MCP, o que significa que os usuários podem continuar executando versões desatualizadas localmente com vulnerabilidades de segurança ou recursos ausentes. Isso pode complicar o cumprimento dos requisitos de conformidade. Algumas ferramentas IDEs e CLI, como o Kiro, permitem que as organizações gerenciem e controlem quais ferramentas MCP estão disponíveis, garantindo consistência e segurança entre as equipes.

Hospedagem remota

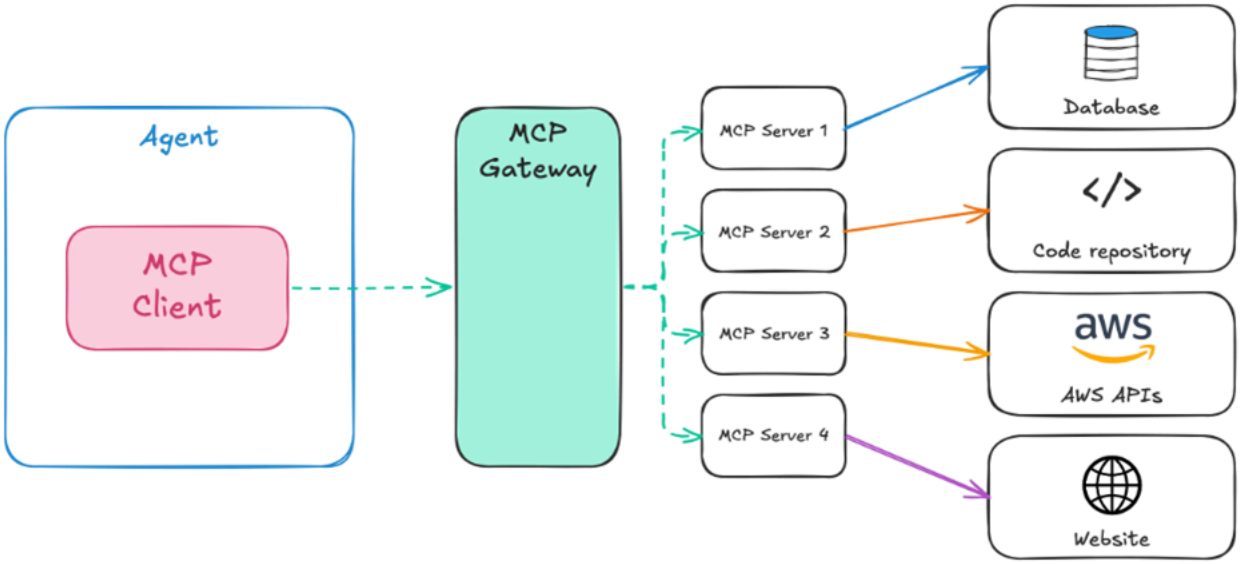
A segunda opção é hospedar servidores MCP remotos que são acessados por HTTP ou HTTPS. Isso fornece acesso a qualquer cliente conectado à rede. O uso da hospedagem remota permite que você controle centralmente o acesso aos recursos e capacidades do MCP, implemente autenticação e autorização e controle o controle de versão e as atualizações da lógica do servidor MCP. A hospedagem remota ainda exige o uso de um registro MCP para que os usuários finais possam descobrir os servidores MCP que desejam usar com seu agente. A imagem a seguir mostra a abordagem de hospedagem remota.



Do ponto de vista do desenvolvimento do agente, a experiência é semelhante, independentemente de o servidor MCP ser local ou remoto. A mudança mais significativa é implementar a autenticação e a autorização, incluindo o acesso do agente ao servidor MCP e o acesso do servidor aos recursos externos. As implementações remotas do servidor MCP devem ser cuidadosamente planejadas para considerar o acesso multilocatário e o gerenciamento de privilégios. O capítulo sobre estratégia de governança do MCP contém mais informações sobre considerações de autenticação e autorização.

Gateway MCP

A opção final é usar um gateway MCP. Os gateways MCP atuam como um proxy centralizado entre clientes e servidores MCP e orquestram o acesso aos servidores MCP registrados. Sem um gateway, cada agente precisa registrar todos os servidores MCP remotos que talvez queira usar. Um gateway permite que o agente se conecte a um único endpoint que gerencia a autenticação, a autorização, o roteamento e a tradução do protocolo. Novos servidores e ferramentas MCP podem ser adicionados dinamicamente e disponibilizados imediatamente para o agente. A imagem a seguir mostra a abordagem do gateway MCP.



Algumas soluções de gateway, como o Docker MCP Gateway, também gerenciam o ciclo de vida dos servidores MCP, lançando servidores sob demanda conforme necessário. Os gateways MCP, como o Amazon Bedrock AgentCore Gateway, também podem ajudar a gerenciar a descoberta de ferramentas fornecendo recursos de pesquisa semântica nativa. Isso fornece aos agentes um único endpoint para se conectar a um cliente MCP e ajuda a otimizar o uso da janela de contexto. O resultado são agentes simples que podem escolher e usar as ferramentas MCP de forma eficaz. No entanto, ele tem desafios relacionados à identidade semelhantes aos da abordagem de servidor MCP remoto.

Práticas recomendadas para hospedar servidores MCP

- O espectro de opções de hospedagem não é único para todos. Grande parte do uso de servidores MCP atualmente é local.
- Ao começar a usar servidores MCP remotos, sua principal consideração é a autenticação e a autorização consistentes para o servidor MCP e como o servidor MCP executa a autenticação e a autorização dos recursos downstream.
- Os gateways MCP simplificam a conectividade, a autenticação e a autorização para hospedar vários servidores MCP remotos. Eles também fornecem recursos para melhorar o gerenciamento de janelas de contexto pesquisando as ferramentas aplicáveis.

Estratégia de governança do MCP

A outra capacidade crítica que o MCP oferece às organizações é o suporte à governança centralizada. Sua estratégia de governança de MCP deve abordar a autenticação e a autorização tanto para os servidores MCP quanto para os recursos que eles acessam. Também deve abordar a limitação de taxas para proteger os recursos posteriores, as métricas operacionais para monitorar o uso e o desempenho das ferramentas e gerenciar implantações e distribuição de servidores MCP.

Autenticação e autorização

Uma das partes mais importantes de sua estratégia de autenticação e autorização é gerenciar o acesso a recursos downstream dos servidores MCP. Quando um usuário chama um agente, a autenticação e a autorização são realizadas para garantir que o usuário tenha permissões para chamar o agente. Em seguida, o agente orquestra a chamada de ferramentas específicas nos servidores MCP. Você precisa decidir como autorizar o acesso por ferramenta.

Uma opção é a machine-to-machine autorização, em que o consentimento ou a interação do usuário não são necessários. Por exemplo, uma invocação de agente baseada em tempo usa um servidor MCP para coletar registros de um aplicativo e analisá-los. Nesse cenário, o agente está pré-autorizado a acessar os dados especificados. A segunda opção é o acesso delegado pelo usuário, em que um usuário fornece seu consentimento para acessar dados e recursos específicos do usuário.

A tabela a seguir mostra os padrões de autenticação e autorização.

Fator	Acesso delegado pelo usuário	Machine-to-machine
Propriedade de dados	Autorização específica do usuário para dados	Dados de todo o sistema ou da organização
Interação com o usuário	O usuário está presente e pode consentir	Sem interação com o usuário
Tempo de operação	Interativo ou em tempo real	Em segundo plano, programado ou em lote
Escopo de permissões	As permissões variam de acordo com o usuário	Permissões consistentes no nível do agente

O acesso delegado pelo usuário requer uma implementação cuidadosa e deve ser desenvolvido com sua equipe de segurança. Os agentes devem ser capazes de avaliar quais ferramentas um LLM selecionou e se elas precisam de autorização adicional. As ferramentas MCP devem incluir descrições para indicar seus requisitos de autenticação e autorização e onde recuperar os tokens de acesso. Os aplicativos cliente devem oferecer suporte a solicitações de autenticação intermediárias, e o cliente MCP deve fornecer as credenciais recuperadas de volta ao agente para cada chamada de ferramenta.

Você deve garantir que as ferramentas MCP sempre tenham seus próprios tokens para acessar recursos externos e que o acesso seja registrado e auditado. As credenciais do usuário não devem ser propagadas por meio de seu sistema agente. Por exemplo, seus servidores MCP não devem usar o mesmo token para acessar dados que foram usados para invocar o agente. As chamadas downstream devem usar tokens com escopo explícito e gerados para fins específicos. Isso ajuda a fornecer proteções adicionais para impedir o acesso não intencional aos dados em nome das ações. Também pode ajudar a evitar que alucinações produzam resultados indesejados. Imagine que um usuário com permissões totais de administrador peça a um agente que clone um banco de dados de produção para uso na pré-produção. Para fazer isso, o usuário só precisa de `CREATE` permissões `READ` e permissões. Digamos que o LLM alucine e acredite que precisa limpar o banco de dados antigo como parte dessa solicitação. Se ele reutilizar as credenciais do usuário, provavelmente será bem-sucedido porque as credenciais originais do usuário têm permissões. `DELETE` Em vez disso, se o servidor MCP usar um token intencionalmente reduzido para a solicitação com apenas `READ` `CREATE` permissões, a tentativa de excluir o banco de dados de produção falhará.

Você pode usar o Amazon Bedrock AgentCore Identity para ajudar a implementar esses padrões. Certifique-se de fazer uma escolha intencional sobre se as permissões para listar e invocar ferramentas hospedadas por um servidor MCP implicam permissão para os recursos externos que o servidor MCP expõe. Esse fluxo de identidade do servidor MCP para o recurso e de volta para o usuário depende do tipo de serviço de autenticação e autorização que está sendo usado. Você deve decidir como isso é tratado em grande escala para seus servidores MCP.

Ao projetar seus padrões de autenticação e autorização, implemente mecanismos de isolamento de token que recuperem diferentes tokens de acesso para cada ferramenta acessada. Não reutilize tokens entre ferramentas e servidores. AgentCore A identidade fornece esse recurso de isolamento de token. Ele gerencia automaticamente os tokens da carga de trabalho (para machine-to-machine autenticação) e os tokens do usuário (para acesso delegado pelo usuário) para garantir a separação adequada e evitar o escalonamento de permissões. Isso é especialmente importante ao incorporar servidores MCP remotos ou gateways MCP.

Melhores práticas para autenticação e autorização de MCP

- Separação de tokens — Não passe tokens portadores dos chamadores para os serviços posteriores. Valide se o campo aud (audiência) corresponde ao servidor que está recebendo o token. A reivindicação do público especifica para qual serviço o token se destina, impedindo a reutilização não autorizada de tokens em diferentes servidores MCP.
- Selecione uma abordagem de acesso — Escolha entre machine-to-machine acesso delegado pelo usuário para cada ferramenta que seus servidores MCP fornecem. Considere agrupar ferramentas no mesmo servidor MCP que usam o mesmo padrão de autenticação.

Controlando a carga

Como em qualquer sistema distribuído, você deve considerar como controlar a carga em sua frota de servidores MCP. Primeiro, você considera se deve implementar a limitação de taxa em seus servidores MCP e onde implementar os limites. Se você optar por não implementar a limitação de taxa, você repassa qualquer limitação de taxa executada por recursos posteriores. Muitos sistemas optam pelo limite de tarifas com base nos atributos da solicitação, como ID do usuário ou da conta. Valide se as solicitações enviadas aos serviços downstream têm esses mesmos atributos para que vários usuários não sejam afetados pela carga gerada por outro usuário.

Se você optar por implementar a limitação de taxa, a abordagem recomendada é implementar a limitação de taxa primária no nível do servidor MCP, com serviços de back-end fornecendo proteção secundária e agentes adaptando seu comportamento com base no feedback do limite de taxa. Considere se os limites de taxa são por servidor MCP ou por ferramenta. Os limites de taxa por servidor MCP ajudam a proteger sua frota e serviços de servidores MCP em um ambiente multilocatário. No entanto, isso pode ser muito grosseiro. Os limites de taxa por ferramenta foram projetados para evitar a sobrecarga de recursos posteriores que podem não se limitar suficientemente. Se uma ferramenta chamar várias APIs, você deve definir o limite de taxa para se alinhar à taxa mais baixa permitida por elas APIs.

A transmissão de informações de limite de taxa em cabeçalhos HTTP também pode ser uma métrica útil para usuários e sistemas automatizados para ajudar a gerenciar sua própria taxa de solicitações e estratégias de repetição. Por exemplo, você pode enviar esses cabeçalhos de volta para o agente a partir do seu servidor MCP, conforme mostrado no exemplo a seguir:

```
X-RateLimit-Limit: 100
X-RateLimit-Remaining: 45
```

```
X-RateLimit-Reset: 1640995200
```

Além disso, considere a redução de carga para proteger o serviço geral quando nenhum cliente estiver excedendo um limite de taxa, mas a carga estiver afetando o desempenho do sistema.

Melhores práticas para controlar a carga

- Escolha uma abordagem de limitação de taxa — Planeje limitar a taxa de usuários individuais com base no uso de recursos posteriores ou no uso de seu servidor e ferramentas MCP.
- Considere a redução de carga — proteja sua frota de servidores MCP da sobrecarga geral que não é causada por um único ou poucos clientes.

Métricas operacionais

As principais métricas a serem capturadas para implementações de MCP devem se concentrar na experiência do cliente que elas oferecem. Essas métricas geralmente incluem uso de tokens, precisão na seleção de ferramentas, número de ferramentas registradas com o agente e latência da ferramenta. Por exemplo, monitorar os tokens de saída retornados por cada ferramenta permite que você defina alarmes quando as ferramentas excedem um limite para o uso da janela de contexto. Quando uma ferramenta excede esse limite, talvez você queira revisar o comportamento da ferramenta. Isso também está relacionado à estratégia de design da ferramenta MCP. As métricas de precisão da seleção de ferramentas indicam o quão bem os agentes escolhem as ferramentas apropriadas para determinadas tarefas, enquanto a velocidade de execução e as taxas de sucesso destacam gargalos de desempenho e problemas de confiabilidade.

Por exemplo, para avaliar as métricas de seleção e precisão do uso de ferramentas, AWS as equipes criaram conjuntos de dados dourados para testes de regressão. Os conjuntos de dados foram gerados sinteticamente usando registros históricos LLMs de invocação da API em consultas de usuários. Usando as métricas predefinidas de seleção e uso de ferramentas (como precisão da seleção de ferramentas, precisão dos parâmetros da ferramenta e precisão das chamadas de funções em vários turnos), AWS as equipes poderiam avaliar objetivamente a capacidade do agente de IA de identificar corretamente as ferramentas apropriadas, preencher seus parâmetros com valores precisos e manter sequências coerentes de invocação de ferramentas em turnos de conversação.

Medir métricas sobre o número de ferramentas registradas com um agente pode ajudá-lo a identificar possíveis desafios de gerenciamento de janelas de contexto, bem como mudanças nas ferramentas

disponíveis apresentadas pelos servidores MCP. Você deve revisar regularmente as métricas operacionais que indicam a experiência do usuário com o servidor e as ferramentas MCP.

Colaboradores

Autoria

- Alex Torres, arquiteto de soluções sênior, AWS
- Saikat Gomes, gerente sênior de soluções para clientes, AWS
- Mike Haken, arquiteto sênior de soluções, AWS
- Sereja Das, engenheira principal, AWS

Análise

- Ted Swinyar, gerente de arquitetura de soluções, AWS
- Raju Patil, cientista de dados sênior, AWS

Redação técnica

- Lilly AbouHarb, redatora técnica sênior, AWS

Histórico do documento

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

Alteração	Descrição	Data
Publicação inicial	—	16 de março de 2026

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.