Operacionalizando a IA agente em AWS

AWS Orientação prescritiva

AWS Orientação prescritiva: Operacionalizando a IA agente em AWS

Table of Contents

Operacionalizando a IA agente em AWS

Aaron Sempf, Brad Ryan, Bhargs Srivathsan e Akhil Bhaskar, da Amazon Web Services

Agosto de 2025 (histórico do documento)

A IA agente não é um recurso, é um novo paradigma operacional. Organizações que investem em arquitetura disciplinada, estruturas de confiança e modelos de implantação alinhados aos negócios liderarão a próxima geração de empresas adaptáveis e inteligentes.

A IA agente representa a convergência de agentes de software autônomos e IA generativa. Ele combina a tomada de decisão e o comportamento direcionado a objetivos dos agentes com a compreensão da linguagem e as capacidades de geração de grandes modelos de linguagem (). LLMs Esses agentes podem raciocinar, agir, se adaptar e colaborar em ambientes corporativos dinâmicos. Para operacionalizar esse potencial, as empresas devem mudar sua mentalidade da implantação de modelos para a infraestrutura de agentes.

Este guia fornece uma estratégia organizacional para transformar a IA agente de experimentos isolados em uma infraestrutura de geração de valor em escala corporativa. Ele pode ajudar você a incorporar agentes inteligentes em todos os fluxos de trabalho com governança, escalabilidade e alinhamento comercial.

Principais áreas de foco e recomendações

Este guia se concentra nas seguintes áreas fundamentais ao operacionalizar a IA agente. Recomendações organizacionais e comerciais são fornecidas para cada área de foco:

- Área de foco 1: Esclareça a intenção e o escopo do agente— Alinhe os agentes às prioridades de negócios e aos gargalos cognitivos. Trate os agentes como colegas de equipe digitais, não apenas como ferramentas.
- Área de foco 2: Design para composição e colaboração— adote sistemas multiagentes com arquitetura modular, protocolos semânticos e delegação dinâmica por meio de agentes árbitros.
- Área de foco 3: Arquiteto para multilocação e controle— crie uma infraestrutura escalável e voltada para inquilinos com serviços de agentes compartilhados, governança centralizada e acesso baseado em funções.
- Área de foco 4: Construir confiança por meio de identidade, barreiras e observabilidade— imponha a rastreabilidade, os controles de tempo de execução e a explicabilidade para ganhar a confiança das partes interessadas.

- Área de foco 5: Gerenciar o ciclo de vida— estabeleça pipelines de integração contínua e implantação contínua (CI/CD), controle de versão imediato, telemetria e reciclagem contínua para apoiar o desempenho e a eficiência da IA agente.
- Área de foco 6: Alinhar modelos de agentes com modelos de negócios— monetize os recursos dos agentes por meio de modelos baseados no uso, métricas internas de ROI e ofertas comerciais.

Você pode usar as recomendações deste guia para preparar sua empresa para a IA agente em grande escala. Ele descreve como as organizações devem se reestruturar em torno da IA agente, incluindo a criação de equipes DevOps de agentes (AgentOps), sistemas interoperáveis e estratégias de gerenciamento de mudanças que ampliem a adoção. Ele enfatiza o pensamento de decisão em primeiro lugar e o alinhamento com o AWS Well-Architected Framework.

Público-alvo

Este guia é destinado a arquitetos corporativos, líderes de AI/ML engenharia e estrategistas de transformação digital que estão projetando e escalando sistemas agentes, incorporando a IA nos principais fluxos de trabalho de negócios e operacionalizando LLMs agentes autônomos em ambientes de produção. Para entender os conceitos e recomendações deste guia, você deve estar familiarizado com as arquiteturas modernas nativas da nuvem e os sistemas distribuídos, os grandes modelos de linguagem, os recursos do modelo básico e os princípios de governança de IA e engenharia de DevOps plataforma.

Objetivos

Ao implementar as recomendações deste guia, sua organização pode alcançar os seguintes resultados comerciais:

- Tomada de decisão acelerada e execução do fluxo de trabalho por meio de agentes autônomos e orientados a metas que reduzem os gargalos humanos e a carga cognitiva.
- Implantações escaláveis e econômicas de recursos inteligentes em todas as unidades de negócios, por meio de plataformas de agentes reutilizáveis e multilocatários.
- Maior resiliência, confiança e governança nos sistemas de IA, o que permite uma adoção segura em ambientes regulamentados, de missão crítica ou voltados para o cliente.

Sobre esta série de conteúdo

Este guia faz parte de uma série sobre IA agente em AWS. Para obter mais informações e ver os outros guias desta série, consulte Agentic AI no site da AWS Prescriptive Guidance.

Fundamentos estratégicos para IA agêntica

Os sistemas agentes não são novos. Agentes de software, incluindo automação robótica de processos (RPA) e mecanismos de decisão, existem há décadas. Mas eles eram simples e determinísticos, projetados para seguir regras predefinidas e lógica simbólica para executar tarefas repetitivas e de baixa variação. Com o surgimento da IA generativa, o jogo mudou. Modelos de linguagem grandes (LLMs) agora podem interpretar entradas complexas, gerar respostas dinamicamente e sintetizar rapidamente o conhecimento. Agora você pode escalar a agência sem uma lógica frágil ou codificada. Agora, os agentes podem raciocinar, tomar decisões, invocar ferramentas, se adaptar ao contexto e se coordenar com outros agentes em todos os fluxos de trabalho. Eles podem operar de forma autônoma em direção a metas, manter a memória e refletir sobre os resultados.

No entanto, a capacidade bruta não é suficiente. Inteligência sem integração gera novidades, não impacto. Para extrair valor do poder LLMs, as empresas devem passar de experimentos isolados para ecossistemas projetados. Os agentes devem ser tratados como serviços de nível de produção que operam sob a mesma disciplina de qualquer sistema corporativo. Isso inclui governança, observabilidade, modelos de identidade seguros e gerenciamento do ciclo de vida. Eles também devem resultar em resultados comerciais reais, não em potencial especulativo. Esses sistemas devem ser arquitetados com limites claros para a tomada de decisões e tolerância a falhas. É importante incorporar mecanismos de recuperação automatizados, monitoramento de desempenho em tempo real e gerenciamento escalável de recursos. Isso ajuda você a lidar com a natureza dinâmica e não determinística das interações dos agentes, mantendo níveis de serviço consistentes em todos os fluxos de trabalho corporativos.

Em um nível fundamental, as empresas devem repensar como a inteligência é incorporada à estrutura das operações. Os agentes devem ser projetados para se integrarem aos sistemas principais, cumprirem as políticas corporativas e fornecerem valor mensurável. Eles precisam operar em grande escala, em todos os departamentos, domínios e contextos de usuário. Em última análise, operacionalizar a IA agente tem a ver com o uso; é a diferença entre implantar uma IA que executa tarefas isoladas e implantar agentes que evoluem seu modelo de negócios.

A IA agente representa uma nova filosofia operacional que exige uma mudança fundamental na forma como abordamos sistemas, processos e pessoas para escalar a inteligência em toda a organização. Os agentes se tornam ativos estratégicos que amplificam as capacidades humanas. Ao integrar a IA agente em suas operações, as organizações podem descobrir insights que geram valor comercial, aumentam as capacidades humanas e otimizam fluxos de trabalho complexos.

Áreas de foco estratégico para IA agente

Para passar dos primeiros protótipos para sistemas de nível de produção e geração de valor, as equipes precisam de uma estratégia coerente que combine arquitetura, processo e pensamento de produto.

Muitas organizações ainda abordam a IA com uma mentalidade centrada na ferramenta ou no modelo. A IA generativa amplificou a experimentação, mas muitas vezes sem um alinhamento claro com a estratégia de negócios ou resultados mensuráveis. Sem um papel estratégico definido, os agentes correm o risco de se tornar novos experimentos que drenam recursos em vez de oferecer valor escalável. Para estabelecer o papel estratégico da IA agente, as organizações devem começar com as prioridades de negócios. Identifique áreas de sobrecarga cognitiva, gargalos de decisão ou fluxos de trabalho fragmentados em que a autonomia pode proporcionar alívio. Use declarações de problemas específicos do domínio para definir as responsabilidades do agente. Trate os agentes como colegas de equipe digitais, não como ferramentas, capazes de raciocinar, delegar e se adaptar.

As ciências da decisão são a disciplina que combina ciência de dados, análise e modelagem comportamental para melhorar a tomada de decisões. Ele deve ser integrado logo no início do processo de arquitetura do agente para alinhar o design aos resultados comerciais. Ao identificar padrões de decisão, simular compensações e quantificar o impacto no valor, as ciências da decisão podem ajudá-lo a identificar onde a autonomia do agente pode oferecer o maior valor. As ciências da decisão podem acelerar decisões, reduzir erros e permitir adaptações em tempo real. Essa base baseada em dados baseia o design do agente em insights mensuráveis e permite uma maior integração com as tecnologias corporativas existentes, como mecanismos de regras, plataformas de análise e modelos preditivos.

Para ajudar a estabelecer o papel estratégico dos agentes, esta seção apresenta as áreas de foco fundamentais que formam a espinha dorsal da operacionalização da IA agente. Cada um é mapeado para um trabalho principal a ser realizado da perspectiva de um líder técnico, arquiteto ou proprietário do produto, responsável pela forma como os agentes são concebidos e projetados. Essas áreas de foco não são etapas sequenciais. Vale a pena revisitá-las em todo o ciclo de vida do sistema para cultivar ecossistemas de agentes resilientes, escaláveis e monetizáveis.

Esta seção contém as seguintes áreas de foco:

- Área de foco 1: Esclareça a intenção e o escopo do agente
- Área de foco 2: Design para composição e colaboração

Área de foco 1: Esclareça a intenção e o escopo do agente

Job to be done: “Ajude-me a garantir que cada agente resolva um problema real com limites claros, não apenas com uma demonstração bacana.”

A IA agente não se trata apenas de desenvolver capacidades. Trata-se de resolver o problema certo, da maneira certa, para obter o resultado certo. Isso começa com uma clareza total sobre a intenção da solução de IA agente.

Estratégia

Muitas vezes, as organizações começam com o que o modelo pode fazer (como ligar APIs, responder perguntas ou gerar resumos) e adaptam um caso de uso em torno dele. Isso leva a um aumento no escopo, à integração deficiente e a agentes que são tecnicamente impressionantes, mas operacionalmente inúteis. Em vez disso, comece definindo a função do agente por meio de perguntas específicas, como as seguintes:

- Por qual resultado específico o agente é responsável?
- De quem está agindo em nome?
- Quem se beneficia?
- Onde começa e termina a autonomia do agente?
- O que acontece quando ele falha?

Um agente bem definido tem um trabalho claro, responsabilidades definidas e critérios de sucesso mensuráveis. Não pense no agente como um assistente ou um chatbot. Em vez disso, dê a ele um título profissional. Pense nisso como um agente de sucesso do cliente, um manipulador de devoluções de produtos ou um monitor de conformidade.

Ao engajar partes interessadas ou clientes, enfatize a escalabilidade e a adaptabilidade dos sistemas de IA agentes. Esses agentes evoluem com a empresa, melhorando continuamente por meio

de aprendizado e feedback. Para reduzir a resistência e acelerar a adoção, destaque como as ferramentas agênticas são projetadas pensando na empatia dos funcionários. Eles fornecem transparência, controle e mecanismos opcionais de substituição que criam confiança. Em vez de substituir pessoas, os agentes aumentam a capacidade humana e a tomada de decisões, ajudando os funcionários a se manterem informados e a se concentrarem em tarefas de alto valor.

A chave para uma implementação bem-sucedida é alinhar a IA agente com resultados comerciais específicos e de alto impacto. Incentive equipes e parceiros a começarem com projetos piloto focados que resolvam pontos problemáticos visíveis. Ganhos rápidos geram retorno sobre o investimento (ROI) mensurável, criam adesão interna e criam impulso para uma adoção mais ampla.

Para orientar a adoção e a maturidade, as organizações podem estruturar o design dos agentes de acordo com um modelo evolutivo. A autonomia, a complexidade e o impacto nos negócios do agente aumentam progressivamente. A seguir estão os estágios desse modelo:

- Agentes observadores revelam insights a partir do ruído. Um exemplo é um agente de opinião do mercado que monitora a percepção da marca em todos os canais digitais.
- Agentes assistentes apoiam a tomada de decisões humanas. Um exemplo é um agente consultivo de negócios que sintetiza dados da concorrência e condições de mercado para equipes de vendas.
- Agentes autônomos agem de forma independente dentro de limites definidos. Um exemplo é um agente de alocação de recursos que ajusta dinamicamente a infraestrutura de nuvem com base na demanda.
- Os agentes do Orchestrator coordenam os fluxos de trabalho de vários agentes. Um exemplo é um agente de otimização da cadeia de suprimentos que gerencia as interações entre agentes de inventário, logística e previsão.
- Agentes inovadores geram novas possibilidades estratégicas. Um exemplo é um agente de inovação de modelos de negócios que analisa as tendências do mercado e recomenda novos fluxos de receita.

Enquadrar os agentes em torno desses resultados estratégicos e níveis de maturidade aumenta o foco, acelera a adoção e aumenta a confiança das partes interessadas.

Para apoiar o alinhamento nessa área de foco Serviços da AWS, como o Amazon Quick, pode visualizar os principais indicadores de desempenho (KPIs) vinculados a resultados orientados por agentes. Você pode usar CloudWatch a Amazon para monitorar o comportamento do agente, as métricas de desempenho e a integridade do sistema quase em tempo real. Use o feedback

operacional para ajustar as interações dos agentes e o uso de recursos. AWS CloudTrailpode fornecer visibilidade da atividade do agente e dos padrões de integração durante as fases iniciais de experimentação e refinamento.

O valor comercial da definição da intenção e do escopo

A adoção da IA agente representa uma mudança fundamental na forma como as organizações abordam a transformação digital e a excelência operacional. Não se trata apenas de automação. Trata-se de permitir uma autonomia inteligente que acelera a tomada de decisões e a realização de valor.

Os principais fatores de negócios incluem o seguinte:

- Vantagem competitiva — Os primeiros usuários obtêm vantagem estratégica por meio de insights mais rápidos, melhores serviços e operações adaptáveis.
- Aprimoramento da experiência do cliente — Os agentes oferecem suporte em tempo real, personalizado e sempre ativo, que aumenta a satisfação e a fidelidade.
- Eficiência operacional — a IA agente reduz significativamente a carga cognitiva humana ao automatizar tarefas de decisão complexas e repetitivas. Isso permite que a equipe se concentre em atividades de maior valor e possa reduzir custos.

Os casos de uso do mundo real em todos os setores incluem o seguinte:

- Serviços financeiros — os agentes de IA podem fornecer aconselhamento financeiro personalizado e detectar fraudes.
- Assistência médica — Agentes de triagem e plano de tratamento podem melhorar a produtividade clínica.
- Varejo — os agentes podem atuar como assistentes de compras inteligentes ou otimizar o estoque em tempo real.
- Fabricação — Os agentes podem realizar manutenção preditiva ou coordenar cadeias de suprimentos.

Área de foco 2: Design para composição e colaboração

Job a ser feito: “Deixe-me criar agentes como eu crio serviços: modulares e testáveis, para que possam ser compostos e orquestrados conforme necessário.”

Muitos esforços de IA começam como pilotos monolíticos e centrados em modelos. Eles são úteis, mas são difíceis de escalar em vários domínios ou de se adaptar a problemas complexos. Valorize os compostos quando esses agentes são projetados para interoperar. Em tecnologia, a capacidade de composição é o ato de combinar componentes modulares para criar uma solução flexível e escalável que possa se adaptar às mudanças. Sem capacidade de composição, a inteligência fica bloqueada em fluxos de trabalho específicos. Além disso, a colaboração de agentes introduz complexidades de orquestração, gerenciamento de estado e negociação de protocolos que as equipes de automação tradicionais talvez não estejam preparadas para lidar.

Estratégia

Adote o paradigma multiagente. Agentes de modelo, como departamentos organizacionais: modulares, especializados e interoperáveis. Defina interfaces claras, formatos de contexto compartilhado e protocolos de comunicação padrão, como Model Context Protocol (MCP) ou Agent2Agent (A2A). Adote padrões de orquestração de vários agentes, como enxame, gráfico ou coordenação hierárquica. Esses padrões ajudam os agentes a descobrir recursos e solicitar serviços uns dos outros de forma dinâmica, seja em fluxos de trabalho paralelos, sequenciais ou orientados por consenso, dependendo da estrutura da tarefa e do nível de confiança.

Para promover uma colaboração escalável e governada, use um agente árbitro. Esse tipo de agente é uma autoridade neutra que facilita a delegação de tarefas com base em capacidades conhecidas e estratégias alternativas. Embora não seja um controlador centralizado, um agente árbitro desempenha um papel fundamental na confiança e na conformidade. Ele garante que tarefas confidenciais ou regulamentadas sejam encaminhadas somente para agentes que atendam aos requisitos de identidade e política. Ele atua como um guardião para fluxos de trabalho vinculados a políticas. Ele impõe o isolamento e permite uma delegação explicável. Crucialmente, um agente árbitro não é um gargalo; ele coexiste com agentes autocoordenados que operam de maneira horizontal. peer-to-peer Esses agentes delegam subtarefas, compartilham contexto e resolvem dependências diretamente.

Esse modelo híbrido suporta tanto a atribuição determinística (por meio do agente árbitro) quanto a colaboração emergente. Ele combina estrutura com flexibilidade. Dentro dessa arquitetura, os agentes podem ser classificados nas seguintes funções especializadas:

- Agentes de decisão, como aplicadores de políticas, alocadores de recursos e avaliadores de risco
- Agentes de conhecimento, como agregadores de contexto, reconhecedores de padrões e detectores de anomalias

- Agentes de execução, como executores de tarefas, controladores de qualidade e gerentes de integração

Para coordenar de forma eficaz, os sistemas multiagentes devem oferecer suporte a protocolos de interação robustos para gerenciamento de estados, recuperação de falhas e resolução de conflitos. Isso promove estabilidade e responsabilidade, mesmo quando os agentes operam de forma independente.

Estabeleça regras claras para escalabilidade, como instanciação de agentes com base em carga, alocação de recursos com reconhecimento de contexto e descoberta e registro automatizados de recursos. Essas medidas ajudam o sistema a crescer dinamicamente em resposta à demanda ou à complexidade.

Projete agentes para serem ready-to-use módulos dentro de um substrato de mensagens distribuído. Por exemplo, você pode usar a Amazon EventBridge com A2A ou MCP em vez de serviços em silos. Adote versões, CI/CD pipelines e modelos de agentes para apoiar a estabilidade do sistema e, ao mesmo tempo, acelerar a adoção interna e a evolução do ciclo de vida. Incentive a reutilização e a padronização do código para reduzir o atrito de integração e promover um ecossistema resiliente.

A colaboração é um multiplicador de forças. Ele libera escala, especialização e resiliência em ambientes multiagentes. Para apoiar essa colaboração dinâmica, as organizações devem arquitetar um plano de controle leve para a coordenação dos agentes. Esse plano de controle inclui o seguinte:

- Registros de recursos que definem o que cada agente pode fazer e oferecem suporte a metadados versionados para descoberta entre pares
- Lógica de arbitragem de tarefas que usa agentes árbitros ou supervisores para rotear tarefas com base no contexto, disponibilidade e política
- Rastreamento do ciclo de vida e do estado que permite o contexto de decisão em tempo real e transferências seguras

Os planos de controle garantem que os sistemas multiagentes permaneçam extensíveis, alinhados às políticas e tolerantes a falhas, sem centralizar a autoridade ou retardar as operações.

No entanto, ambientes multiagentes também trazem desafios operacionais. Manter o contexto entre as interações dos agentes, gerenciar o estado compartilhado e coordenar ações pode gerar complexidade e custo. Os custos podem aumentar se você usar esses LLMs tokens de consumo durante a comunicação entre agentes. Esses custos devem ser ponderados em relação aos benefícios comerciais combinados da autonomia inteligente em grande escala.

Para enfrentar esses desafios, considere plataformas agentes que abstraiam as principais preocupações, como as seguintes:

- Protocolos de comunicação padronizados e formatos semânticos
- Lógica de orquestração integrada e roteamento dinâmico
- Contexto compartilhado e gerenciamento de memória entre agentes
- Tratamento de falhas e degradação suave durante falhas

Para equipes que adotam estratégias multiagentes, a melhor abordagem é começar aos poucos e projetar para escalar. Comece com soluções específicas de agente único que resolvam problemas reais. Em seguida, componha progressivamente esses agentes em um sistema cooperativo em que cada um possa descobrir, coordenar e delegar com base em metas compartilhadas e no contexto de todo o sistema.

É importante ressaltar que o tratamento robusto de erros e a degradação suave devem ser os principais princípios do projeto. Os sistemas multiagentes devem ser capazes de continuar fluxos de trabalho parciais ou iniciar a lógica de backup quando os agentes não estão disponíveis ou falham. Isso promove confiabilidade sem acoplamento rígido.

Serviços da AWS oferecem recursos robustos para dar suporte a essa arquitetura em grande escala. A Amazon EventBridge e a EventBridge Pipes fornecem o backbone estruturado e orientado por eventos para mensagens com vários agentes. Para gerenciar o comportamento modular, AWS AppConfigpermite alternar configurações seguras e dinâmicas entre instâncias do agente. Para oferecer suporte ao contexto compartilhado e ao gerenciamento de memória, use o Amazon DynamoDB para uma persistência de estado leve e com reconhecimento de inquilinos e uma rápida recuperação de contexto entre agentes. Você pode usar o Amazon Simple Storage Service (Amazon S3) para armazenar históricos de prompts estruturados, artefatos compartilhados ou saídas geradas por agentes. Para fluxos de trabalho mais complexos que exigem coordenação estável, é AWS Step Functionspossível orquestrar processos de longa execução com pontos de verificação e lógica de recuperação de erros. Juntos, esses serviços ajudam você a criar sistemas multiagentes compostos, resilientes e semanticamente conectados que se adaptam às demandas corporativas.

Valor comercial dos sistemas multiagentes

Embora muitas organizações iniciem sua jornada de IA com soluções de agente único, todo o potencial da IA agente é liberado por meio de sistemas multiagentes escaláveis. Esses sistemas são

essenciais para resolver problemas complexos e distribuídos e criar ecossistemas de IA robustos e flexíveis que evoluem de acordo com as necessidades dos negócios.

Os principais benefícios comerciais dos sistemas multiagentes incluem o seguinte:

- Escalabilidade — As tarefas e cargas de trabalho podem ser distribuídas entre agentes especializados para aumentar a capacidade e o desempenho.
- Flexibilidade — os agentes podem ser adicionados, substituídos ou modificados com o mínimo de interrupção, permitindo agilidade em ambientes dinâmicos.
- Resiliência — a estabilidade do sistema é preservada mesmo quando agentes individuais falham, graças às funções redundantes e ao failover inteligente.
- Especialização — Agentes específicos realizam tarefas com maior eficiência e precisão.
- Eficiência de custos — componentes de agentes reutilizáveis aceleram o desenvolvimento e reduzem o custo da implantação de novos recursos.

Embora os sistemas multiagentes exijam um planejamento mais antecipado, eles oferecem agilidade, velocidade e capacidade de inovação a longo prazo. As empresas que investem em arquiteturas flexíveis de colaboração de agentes estão posicionadas para implantar novos recursos de IA rapidamente, adaptar-se às demandas em constante mudança e liderar em um cenário competitivo cada vez mais orientado por agentes.

Área de foco 3: Arquiteto para multilocação e controle

Trabalho a ser feito: “Ajude-me a escalar o uso de agentes em vários clientes sem perder o controle, a responsabilidade ou a visibilidade”.

Os primeiros protótipos são bons para provar valor isoladamente, mas a maioria das empresas precisa oferecer suporte simultâneo a vários clientes, departamentos ou fluxos de trabalho. Isso significa que cada agente deve operar dentro de limites claramente definidos de políticas, dados e identidade. Sem multilocação, as operações se tornam frágeis e caras, e a governança se torna uma colcha de retalhos.

Estratégia

Siga os princípios das arquiteturas de software como serviço (SaaS). Por exemplo, design para isolamento de inquilinos, aplicação de políticas e controle de recursos. Arquitete agentes e plataformas de orquestração com memória, configuração e identidade sensíveis ao inquilino.

Para impor limites, use marcação, controle de acesso baseado em função (RBAC) e escopo de gerenciamento de identidade e acesso.

Adote uma camada de observabilidade unificada em que a telemetria do agente seja agregada pelo contexto do inquilino. Implemente mecanismos de política centralizados e alternância de recursos baseada em configuração para impor regras de comportamento dinâmico.

Crie a implantação de agentes como um serviço. Permita que equipes internas ou clientes consumam os recursos dos agentes de forma escalável e APIs governada. AWS fornece uma base sólida para esses padrões. Você pode usar o Amazon Cognito para gerenciar a identidade do usuário e do inquilino AWS Organizationse as políticas de controle de serviços (SCPs) para governança entre contas e AWS Resource Access Manager (AWS RAM) para compartilhar recursos com segurança. Além disso, AWS AppConfigpode gerenciar dinamicamente o comportamento do agente por locatário ou ambiente. Esses serviços ajudam a impor limites e políticas e, ao mesmo tempo, oferecem suporte à infraestrutura compartilhada.

Essa transição da implantação estática para o provisionamento dinâmico transforma a IA agente em uma plataforma para toda a empresa.

Valor comercial das plataformas de agentes multilocatários

A multilocação é mais do que uma conveniência arquitetônica — é um acelerador de negócios. À medida que os agentes inteligentes proliferam entre departamentos e equipes, as organizações devem apoiar o crescimento sem duplicar a infraestrutura ou fragmentar a governança.

Os principais benefícios comerciais dos sistemas multilocatários incluem o seguinte:

- Escalabilidade — uma plataforma de agente multilocatário permite que equipes internas, unidades de negócios ou clientes integrem recursos de IA mais rapidamente sem precisar de ambientes personalizados.
- Eficiência de custos — a infraestrutura compartilhada minimiza implantações redundantes, consolida os custos operacionais e simplifica a manutenção em todos os ambientes.
- Governança e redução de riscos — Controles centralizados de políticas, modelos de identidade e observabilidade ajudam os agentes a operar com mais segurança e conformidade em todos os locatários.
- Reutilização do serviço — Para promover a reutilização e reduzir a duplicação, agentes com reconhecimento de inquilinos podem ser oferecidos como serviços internos, como para enriquecimento, conformidade ou resumo.

Exemplos de casos de uso para sistemas multilocatários incluem o seguinte:

- Um agente de conformidade implantado em subsidiárias adapta sua lógica às regulamentações locais por meio da configuração específica do inquilino. Isso elimina a necessidade de criar agentes separados para cada região.
- Um agente interno de automação do fluxo de trabalho atende a vários departamentos com diferentes limites e permissões de dados. Ele mantém o isolamento enquanto acelera o cumprimento das tarefas.

Ao projetar agentes como multi-tenant-aware serviços, as organizações evitam a sobrecarga de iniciativas de IA em silos. Em vez disso, eles promovem uma plataforma de inteligência unificada. Essa arquitetura permite implantação escalável, consistência operacional e melhor ROI. Também facilita a expansão da adoção da IA em toda a empresa.

Área de foco 4: Construir confiança por meio de identidade, barreiras e observabilidade

Trabalho a ser feito: “Dê-me a confiança de que os agentes agirão com segurança e previsibilidade, especialmente quando ninguém estiver assistindo”.

Agentes autônomos desafiam os modelos de controle tradicionais. Sua capacidade de raciocinar e agir de forma independente apresenta riscos se não forem gerenciados adequadamente. Sem restrições claras de propriedade, auditabilidade ou política, eles podem se afastar do comportamento pretendido. Construir confiança organizacional exige mais do que apenas confiabilidade técnica. Ela exige explicabilidade, responsabilidade e consistência.

Estratégia

Crie um sistema de controle que priorize a identidade como a espinha dorsal da autonomia confiável. Cada agente deve operar com uma identidade verificável, permissões com escopo definido e histórico de execução rastreável. Os agentes devem ser incorporados em uma estrutura de confiança zero que inclua vinculação de inquilinos, herança de acesso contextual e aplicação de tempo de execução por meio de proteções e mecanismos de políticas. Isso permite auditar, reverter ou restringir as ações do agente com base nas regras organizacionais e na postura de risco.

Incorpore a fiscalização da confiança em tempo de execução por meio de grades de proteção inteligentes. Isso inclui controles de taxas e limitação com base em padrões comportamentais ou

condições de carga de trabalho, limites de recursos aplicados junto com o auto-scaling e pontuação de decisão para avaliar o risco. Crie acionadores para engajar human-in-the-loop fluxos de trabalho quando os limites forem excedidos.

Todo agente também deve ser transparente e explicável. Incorpore a telemetria estruturada por meio de registros, rastreamentos e resumos de raciocínio para expor a lógica de decisão. Support, trilhas de decisão e rastreamento de impacto. Isso ajuda você a conectar as ações do agente às principais métricas ou resultados. Implemente mecanismos de detecção de desvios que monitorem desvios do comportamento ou das políticas esperados.

Introduza agentes reflexivos que observem continuamente o comportamento dos agentes e os padrões do sistema. Eles devem sinalizar anomalias ou inconsistências em tempo real. Esses agentes contribuem para ciclos de feedback de governança que podem iniciar a revalidação, adaptação ou desativação de recursos.

Estabeleça conselhos de governança que revisem as políticas dos agentes, aprovem mudanças de capacidade e supervisionem os protocolos de resposta a incidentes. A confiança deve ser conquistada, medida e continuamente reforçada.

AWS fornece uma base sólida para a implementação dessa estrutura de confiança:

- AWS Identity and Access Management (IAM) impõe limites de execução e permissão baseados em funções
- Amazon CloudWatch e AWS X-Raysuporte com total visibilidade e rastreabilidade.
- Amazon GuardDuty e AWS Configdetecte anomalias de segurança ou desvios de políticas.

Juntos, esses serviços permitem a imposição de identidade, a segurança do tempo de execução e a governança baseada em confiança em grande escala. Eles podem ajudar a tornar os sistemas autônomos poderosos e confiáveis.

O valor comercial da autonomia confiável

À medida que os agentes se tornam mais autônomos, a confiança se torna um fator essencial para a adoção, a governança e o desempenho operacional da empresa. Estabelecer uma base de identidade, observabilidade e barreiras ajuda as organizações a escalar a IA agente em domínios confidenciais, sem sacrificar a governança ou o controle.

Os principais fatores de negócios incluem o seguinte:

- Garantia de governança — Modelos de identidade fortes, trilhas de auditoria e limites de permissão reduzem o risco de conformidade e apoiam o alinhamento regulatório.
- Continuidade operacional — proteções de tempo de execução e detecção de anomalias ajudam a evitar comportamentos não intencionais e apoiam a autorrecuperação de falhas extremas.
- Confiança das partes interessadas — A explicabilidade das decisões e a telemetria criam confiança com as partes interessadas internas, os gerentes de risco e os auditores externos.
- Resiliência a incidentes — a observabilidade incorporada acelera a análise da causa raiz e o tempo de resposta quando surgem problemas.

Exemplo de casos de uso incluem:

- Nos serviços financeiros, os agentes de detecção de fraudes devem expor seu raciocínio, registrar todas as ações com identidade rastreável e operar sob funções específicas do IAM.
- Na área da saúde, agentes de triagem autônomos devem aplicar verificações de segurança em tempo de execução, passar para a revisão humana quando os limites forem atingidos e fornecer registros completos para supervisão clínica.

Ao incorporar mecanismos de confiança no ciclo de vida do agente, as organizações podem permitir que seus sistemas operem de forma autônoma e com responsabilidade. Essa base reduz o risco e capacita os agentes a agirem em nome da empresa com transparência e integridade.

Em última análise, a autonomia confiável acelera a adoção, dando aos usuários e à liderança a confiança necessária para escalar agentes inteligentes em todas as operações principais.

Área de foco 5: Gerenciar o ciclo de vida

Job a ser feito: “Certifique-se de que minha equipe possa aprimorar os agentes ao longo do tempo, sem caos ou heroísmo”.

Ao contrário dos aplicativos tradicionais que são moldados apenas por código, o comportamento do agente também é moldado por solicitações, memória, ferramentas e contexto de treinamento. Esses fatores variam com o tempo. O desvio diminui a confiabilidade, aumenta os custos e torna a depuração quase impossível. Sem controles do ciclo de vida, os agentes param de agregar valor e começam a acumular riscos.

Estratégia

Estabeleça DevOps para agentes (AgentOps) como uma prática. Integre CI/CD pipelines personalizados para agentes. Use esses pipelines para testar resultados imediatos, validar integrações de ferramentas e traçar o perfil do comportamento de custo-desempenho. Mantenha históricos de versões de solicitações, políticas e interações com modelos.

Use ciclos de feedback dos dados de observabilidade para iniciar o retreinamento, o ajuste imediato ou a retirada do agente. Incorpore mecanismos de reflexão em todo o sistema, como um registro de melhorias, para institucionalizar o aprendizado.

Crie um painel de telemetria de desempenho que mostre a precisão, a latência, o custo e a confiabilidade das decisões. Para simplificar e acelerar o gerenciamento do ciclo de vida usando a AWS infraestrutura, as equipes podem usar kits de ferramentas de agentes. Um exemplo é o SDK Strands Agents, que fornece ferramentas estruturadas para rápido controle de versão, registro de ferramentas e integração de CI/CD com Serviços da AWS, como, e. AWS CodePipelineAWS Cloud Development Kit (AWS CDK)AWS Lambda Além disso, use o Amazon S3 e o Amazon Elastic File System (Amazon EFS) para armazenar artefatos de agentes e dados de treinamento. Use AWS Step Functionspara automatizar fluxos de trabalho complexos de reciclagem ou validação. Você pode usar o Amazon SageMaker AI quando os agentes precisarem de ajustes personalizados de modelos ou ajustes finos de fluxos de trabalho além da orquestração do LLM. A disciplina do ciclo de vida transforma agentes de experimentos em ativos duráveis e em evolução.

Com o tempo, esse sistema de ciclo de vida forma a espinha dorsal da inovação. Ele ajuda você a recompor, treinar novamente e reimplantar recursos com agilidade. Isso transforma a camada de agentes em um sistema vivo, capaz de evoluir em resposta tanto ao feedback quanto à oportunidade.

Valor comercial do gerenciamento do ciclo de vida

O gerenciamento eficaz do ciclo de vida é um fator fundamental para o desempenho e a eficiência de custos dos agentes. Isso garante que os agentes inteligentes continuem a fornecer resultados precisos, confiáveis e alinhados ao valor à medida que evoluem. Por padrão, os agentes não permanecem valiosos. Eles devem evoluir em sincronia com as mudanças nos requisitos de negócios, nos fluxos de trabalho e nos ambientes de dados. Uma AgentOps equipe disciplinada ajuda os agentes a permanecerem precisos, eficientes e alinhados às metas corporativas ao longo do tempo.

Os principais fatores de negócios incluem o seguinte:

- Consistência de desempenho — testes contínuos, validação imediata e reciclagem ajudam os agentes a manter a qualidade das decisões em condições e conjuntos de dados em constante mudança.
- Otimização de custos — A criação de perfil orientada por telemetria identifica ferramentas ineficientes, solicitações de alto valor simbólico ou execuções desnecessárias. Em seguida, você pode ajustar para reduzir os custos operacionais.
- Iteração mais rápida — automação do ciclo de vida com ciclos de desenvolvimento CI/CD acelerados, ajudando as equipes a experimentar, implantar e melhorar os agentes com confiança.
- Redução de riscos — Versionamento imediato, suporte à reversão e mecanismos estruturados de avaliação ajudam a evitar regressões e oferecem suporte ao gerenciamento de mudanças seguro e confiável.

Alguns casos de uso incluem o seguinte:

- Um agente de suporte ao cliente é monitorado quanto à latência, custo do modelo e feedback do usuário. A observabilidade revela um aumento de custo, o que exige o reajuste de seus prompts incorporados e da lógica do modelo alternativo.
- Um agente de resumo de contratos é atualizado com base no feedback das equipes jurídicas. Os prompts versionados são testados em ambientes restritos antes do lançamento da produção, oferecendo suporte à segurança e à qualidade.

Com o gerenciamento estruturado do ciclo de vida, as organizações vão além da manutenção reativa para a melhoria proativa e contínua. Os agentes se tornam ativos digitais adaptáveis que são medidos, refinados e revalidados de acordo com as metas de negócios. Essa prática transforma ecossistemas de agentes em sistemas resilientes, econômicos e de alto desempenho que oferecem valor duradouro enquanto acompanham as mudanças.

Área de foco 6: Alinhar modelos de agentes com modelos de negócios

Job to be done: “Mostre-me o impacto, para que eu possa justificar o investimento contínuo.”

Até mesmo agentes tecnicamente capacitados se tornam passivos se não estiverem vinculados aos resultados comerciais. Os agentes devem oferecer eficiência, monetização ou diferenciação estratégica. No entanto, a maioria das empresas tem dificuldade em definir como os agentes se

encaixam nos modelos de preços, embalagens ou uso. Sem um alinhamento claro com o valor comercial, é difícil justificar a escalabilidade ou até mesmo a manutenção do investimento.

Estratégia

Adote práticas de gerenciamento de produtos. Trate os agentes como serviços monetizáveis com um ROI mensurável. Defina estratégias de preços com base em decisões, sessões ou resultados. Em seguida, agrupe os recursos do agente em ofertas hierárquicas que estejam alinhadas aos segmentos de clientes ou às unidades de negócios internas.

Para promover a sustentabilidade, as organizações devem capturar tanto o valor direto quanto os multiplicadores de crescimento por meio da implantação de agentes. Considere usar as seguintes métricas de ROI para medir o valor imediato:

- Custo por decisão — Compare os custos de processamento do agente com os equivalentes humanos.
- Compressão de tempo — quantifique o valor dos ciclos acelerados, como vendas ou aprovações mais rápidas.
- Redução de erros — Avalie a economia com maior precisão, consistência e conformidade.

Além desses ganhos imediatos, os agentes podem desbloquear as seguintes oportunidades de crescimento a longo prazo:

- Empilhamento de capacidades — combine serviços de agentes para criar soluções verticais específicas de domínio.
- Efeitos de rede — Aumente o valor por meio de ecossistemas multiagentes em que a coordenação aumenta a utilidade.
- Extensão do mercado — gere novos fluxos de receita por meio de serviços externamente consumíveis e habilitados por agentes.

Crie ciclos de feedback a partir de métricas de negócios (como economia de custos, aumento de conversão ou time-to-resolution) para impulsionar a evolução contínua dos agentes. Analise a telemetria de uso e os índices de satisfação do usuário para refinar seu alinhamento de valores e prioridades do roteiro. Ao vincular as capacidades dos agentes diretamente aos modelos de negócios, as organizações se posicionam para capturar valor agregado e sustentável, não apenas resultados técnicos.

Os itens a seguir Serviços da AWS apoiam esse alinhamento fornecendo estruturas robustas de rastreamento e monetização:

- AWS Cost Explorere a Amazon CloudWatch fornecem informações sobre os custos por agente e a eficiência operacional.
- O Amazon API Gateway permite acesso medido, limitação de taxas e preços escalonados para endpoints de agentes.
- AWS Marketplacefornece um canal para agentes editoriais e soluções de agentes como produtos comerciais.

Esses serviços ajudam você a transformar a funcionalidade do agente em ofertas digitais escaláveis e orientadas por valor que se alinham às estratégias de crescimento e monetização da empresa.

Entrega de software em evolução para IA agente

A entrega moderna de software foi moldada por uma suposição simples: que você controla os sistemas que envia. Você define os requisitos, escreve a lógica, testa os resultados esperados e implementa serviços previsíveis. Até mesmo o Agile e DevOps as abordagens ainda se baseiam no princípio de que cada sprint oferece algo determinístico, verificável e, em grande parte, sob a supervisão humana.

A inteligência artificial agêntica derruba essa base. Os sistemas agentes interpretam, raciocinam e adaptam em vez de seguir scripts. O comportamento deles depende do código que você escreve, do contexto em que operam, das entradas que recebem, das ferramentas que podem acessar e das metas que lhes são atribuídas. Em resumo, eles não seguem ordens; eles buscam resultados.

Isso faz com que a entrega tenha menos a ver com controle e mais com alinhamento. Em vez de fornecer instruções, você deve moldar como ele se comporta. Isso significa que o ciclo de vida de desenvolvimento de software tradicional (SDLC) não é mais adequado porque foi projetado para sistemas baseados em lógica e controlados por humanos.

Esta seção contém os seguintes tópicos:

Zonas de intenção para IA agente

Em vez de estágios rígidos, como definir, construir, testar e liberar, precisamos de um modelo que englobe autonomia, incerteza e emergência. Em vez disso, você usa zonas de intenção. A z de intenção define um espaço limitado onde um agente pode operar com autonomia, dentro de restrições. O objetivo é mudar do microgerenciamento de todas as tarefas para o design de ambientes em que os agentes possam agir, aprender e colaborar com segurança. Você especifica o quê (o resultado desejado), o porquê (a intenção) e as barreiras de proteção (as restrições, as políticas e os limites de confiança). Dados esses limites e essas informações, o agente descobre como.

Em vez de uma linha de montagem, pense no ambiente como um espaço aéreo. Você controla quem pode entrar, o que eles podem fazer e aonde podem ir. Mas, uma vez lá dentro, eles ficam livres para navegar conforme necessário. É assim que os sistemas agentes se expandem sem caos.

Essa não é apenas uma mudança filosófica; é uma mudança prática. A saída não determinística dos sistemas baseados em agentes não pode ser totalmente testada por meio de testes unitários. Não pode ser versionado como binários estáticos. Os agentes mudam com o tempo, se adaptam a novos dados e interagem com outros sistemas de maneiras imprevisíveis. Tentar entregá-los usando modelos tradicionais leva a arquiteturas frágeis e não escaláveis. Na pior das hipóteses, isso leva a uma falsa confiança em sistemas que você não pode realmente governar.

Quando as equipes adotam a entrega baseada em intenção, elas ganham duas vantagens:

- Controle onde é mais importante — Eles definem limites em vez de saídas.
- Escalabilidade por meio da delegação — Eles permitem que os agentes lidem com a complexidade que os humanos não conseguem codificar.

É assim que você passa de protótipos isolados para sistemas agentes reais, de nível de produção, que podem agregar valor de forma repetida e confiável.

Evoluindo o ciclo de vida de entrega para IA agente

Para apoiar o comportamento inteligente e adaptativo, o SDLC deve ser reformulado do controle determinístico para a intenção adaptativa. A seguir estão as mudanças necessárias para desenvolver o SDLC tradicional para IA agente:

- O planejamento se torna um design de intenção. As equipes definem metas, restrições e comportamentos esperados dos agentes. As políticas e os critérios de sucesso são estruturados em termos de alinhamento, não de lógica.
- A arquitetura se torna andaime. As equipes se concentram em definir funções, interfaces, barreiras, mecanismos alternativos e observabilidade, em vez de criar scripts para cada caminho de decisão.
- O teste se torna uma avaliação comportamental. Em vez de afirmar resultados específicos, as equipes validam se os agentes permanecem dentro dos limites aceitáveis e cumprem a intenção com informações variadas.
- A implantação se torna uma orquestração contínua. Os sistemas Agentic são implantados com controles de tempo de execução, monitoramento ao vivo e canais de feedback que permitem ajustes em tempo real.
- A iteração se torna feedback e adaptação. Em vez dos ciclos tradicionais de correção de mudança de código, as equipes observam como os agentes evoluem, onde são bem-sucedidos ou quando

mudam. Conforme necessário, as equipes intervêm por meio de restrições atualizadas, reciclagem e adição ou modificação de mecanismos de controle.

As práticas existentes que se concentram em iteração, experimentação e feedback rápido estão na metade do caminho. A mudança para sistemas agentes não é uma rejeição dos princípios ágeis. Na verdade, é uma evolução natural deles. O pensamento ágil enfatiza a adaptabilidade, o feedback e as soluções de trabalho em vez de planos rígidos. Isso se alinha perfeitamente com a natureza dos sistemas agentes, que aprendem, se adaptam e respondem ao contexto em tempo real. Se você já está executando ciclos curtos, validando suposições rapidamente e gerenciando incertezas por meio da entrega contínua, você está bem equipado para liderar essa transição.

Mas existem diferenças importantes. A abordagem ágil tradicional pressupõe que o que está sendo entregue é determinístico. Ele pressupõe que, uma vez construída, a coisa se comportará de forma consistente e previsível, com resultados repetíveis para as mesmas entradas. Essa repetibilidade ajuda você a depurar, testar e iterar com confiança. Os sistemas agentes quebram esse modelo. Eles são probabilísticos, sensíveis ao contexto e capazes de evoluir de forma independente. Isso significa que algumas práticas ágeis se tornam menos úteis, como o rastreamento de velocidade com base na conclusão da história, critérios rígidos de aceitação ou planejamento determinístico de sprint.

Os seguintes aspectos do SDLC tradicional se aplicam à IA agente:

- Desenvolvimento e entrega iterativos
- Feedback do cliente como sinal principal
- Colaboração multifuncional
- Integração e implantação contínuas

Os seguintes aspectos do SDLC tradicional devem evoluir para a IA agente:

- Redefina concluído como alinhado à intenção. Concentre-se em saber se o comportamento do agente satisfaz a meta pretendida dentro das restrições definidas.
- Mude de critérios de aceitação para barreiras comportamentais.
- Expanda a definição de concluído para incluir a prontidão para o tempo de execução, que inclui mecanismos de observabilidade, explicabilidade e feedback que apoiam o aprendizado e a confiança contínuos.

- Priorize ciclos de feedback em tempo real e rastreamento de comportamento em vez do planejamento inicial

A boa notícia é que você não precisa descartar o manual do SDLC. Você só precisa evoluí-lo do gerenciamento de código para a modelagem da conduta. Em sistemas agênticos, o sucesso não depende apenas da execução do software, mas de como ele se comporta.

Preparando equipes para a IA agêntica

A engenharia de software não vai desaparecer. Está evoluindo. O trabalho muda de escrever funções para moldar estruturas e mecanismos de controle para um comportamento inteligente. No mundo da IA agente, construir não é mais a parte mais difícil — gerenciar a emergência é. Para a maioria das equipes de engenharia, a evolução parece mais uma mudança de mentalidade do que um salto técnico. Em vez de perguntar “O que o sistema fará?” a pergunta é “O que a capacitamos a buscar e como saberemos se ela continua em andamento?”

Para as equipes de engenharia, a evolução em direção à IA do agente exige as seguintes mudanças:

- Uma mudança cultural — As equipes devem se sentir confortáveis com a incerteza e a autonomia em sistemas que não controlam totalmente.
- Novas funções — designers de intenções, testadores comportamentais e engenheiros de observabilidade se tornam fundamentais para a entrega.
- Linguagem compartilhada — As equipes precisam de uma compreensão clara e compartilhada das metas, barreiras e sinais de sucesso, assim como antes precisavam de especificações e casos de teste.

À medida que a IA generativa amadurece, veremos mais sistemas agentes interagindo com clientes, produtos e operações. As organizações bem-sucedidas não serão aquelas com os melhores modelos. Serão aqueles que poderão integrar agentes em fluxos de trabalho do mundo real com confiança, controle e velocidade. Isso significa que os modelos de entrega e as equipes de engenharia devem evoluir juntos. As zonas de intenção fornecem a abstração para fazer isso. Eles ajudam você a operacionalizar a autonomia sem abrir mão da responsabilidade. Eles também oferecem uma estrutura compartilhada entre equipes para ajudar a governar sistemas que não podem ser codificados.

Para obter mais informações sobre como preparar equipes para a IA agente, consulte a seção Preparando a empresa para a IA agente em grande escala deste guia.

Preparando a empresa para a IA agente em grande escala

À medida que as áreas de foco descritas neste guia convergem, a IA agente muda de funções isoladas para uma camada de inteligência unificada que pode ser entendida como uma plataforma de recursos. Essa plataforma não executa apenas tarefas. Ele evolui, se adapta e se coordena em todos os domínios. Os agentes se tornam serviços modulares, reutilizáveis e detectáveis que aceleram a inovação, reduzem a carga cognitiva e geram resultados mensuráveis em toda a empresa. Essa visão da plataforma prepara o terreno para uma inteligência escalável incorporada em todo o modelo operacional.

Operacionalizar a IA agente exige mais do que implantar agentes inteligentes. Isso exige uma transformação fundamental na forma como a empresa organiza equipes, projeta processos e governa a tecnologia. Assim como a mudança para a nuvem ou modelos operacionais DevOps redefinidos, a IA agente introduz uma nova era de automação de decisões, aprendizado contínuo e coordenação autônoma. O sucesso depende do alinhamento dos sistemas, das pessoas e dos processos em torno dessa nova filosofia operacional.

Esta seção contém os seguintes tópicos:

- Alinhando equipes e modelos de propriedade
- Gerenciando mudanças e prontidão organizacional
- Arquitetura para interoperabilidade e colaboração
- Construindo a governança em um tecido agêntico
- Adotando uma mentalidade operacional que prioriza a tomada de decisões
- Dimensionamento com propósito e intenção

Alinhando equipes e modelos de propriedade

O primeiro passo em direção à maturidade é o alinhamento interfuncional. As empresas devem estabelecer AgentOps equipes que incluam AI/ML profissionais e especialistas no domínio, como arquitetos de sistemas distribuídos, engenheiros de software, proprietários de produtos, líderes de conformidade e arquitetos de plataformas. Essas equipes são proprietárias conjuntas de todo o ciclo de vida de um agente, desde o design e a implantação até o treinamento e o monitoramento.

O provisionamento e a liberação do agente devem seguir as práticas nativas da nuvem, como usar a AWS Cloud Development Kit (AWS CDK)e AWS CodePipelinepara a infraestrutura como código

e implantação automatizada. Essa estrutura promove a responsabilidade compartilhada e acelera a iteração. Assim como DevOps unifica o desenvolvimento e as operações, AgentOps conecta inteligência com governança e execução.

Para serem eficazes, essas equipes também precisam de uma linguagem compartilhada. As partes interessadas da empresa devem entender o que são os agentes, como eles operam e quais resultados eles geram. O treinamento e a capacitação interna são essenciais. Ao desmistificar os agentes e incorporar esse modelo mental nas conversas diárias, as organizações liberam uma participação mais ampla e uma inovação mais alinhada.

Para acelerar o desenvolvimento e a integração dos agentes que usam Serviços da AWS, as equipes podem adotar estruturas como o Strands Agents SDK, que oferece ferramentas baseadas em CLI para estruturar, configurar e empacotar agentes. O Strands Agents foi projetado para funcionar perfeitamente com a AWS infraestrutura, como Amazon Bedrock, AWS LambdaAmazon EventBridge AWS CDK, the e. AWS CodePipeline Ele permite prototipagem e implantação rápidas, mantendo os padrões de nível de produção.

Mas a estrutura e o ferramental por si só não são suficientes. A escalabilidade da IA agente exige uma preparação cultural, educacional e de liderança deliberada para garantir que a adoção se enraíze em toda a organização.

Gerenciando mudanças e prontidão organizacional

A escalabilidade bem-sucedida da IA agente exige mais do que implantar infraestrutura ou agentes inteligentes. Ela exige uma abordagem estruturada para a mudança organizacional. Isso inclui prontidão cultural, desenvolvimento de habilidades, ciclos de feedback orientados por métricas e alinhamento executivo para garantir que a adoção seja intencional e sustentável.

Promover a evolução cultural

- Posicione os agentes como colegas de equipe, não como substitutos, para reduzir a resistência e criar confiança.
- Comunique-se de forma transparente sobre as capacidades e limitações do agente para definir expectativas realistas.
- Estabeleça protocolos de transferência claros para quando os agentes devem encaminhar as decisões para uma autoridade superior ou delegar partes do processo a um colaborador humano.

Estabeleça uma estrutura de desenvolvimento de habilidades

- Ofereça treinamento baseado em funções personalizado para engenheiros, gerentes de produto, líderes de domínio e agentes de conformidade.
- Crie centros de excelência para compartilhar as melhores práticas, padrões de ferramentas e ativos reutilizáveis.
- Combine especialistas em IA com especialistas do domínio por meio de programas de orientação para preencher as lacunas de conhecimento.

Defina métricas e ciclos de feedback

- Ancore o valor técnico e comercial KPIs ao valor estratégico para avaliar o impacto. Exemplos de valor incluem latência de decisão, precisão da resolução e economia de custos.
- Capture de forma sistemática e contínua o feedback do usuário sobre os pontos de atrito e os desafios de adoção.
- Faça retrospectivas regulares para avaliar o desempenho do agente, as tendências de uso e as oportunidades de melhoria.

Alinhe a liderança do topo

- Obtenha patrocínio executivo vinculando as iniciativas dos agentes aos resultados estratégicos e ao ROI.
- Forme comitês de governança multifuncionais que incluam liderança técnica e comercial.
- Adapte as estratégias de comunicação para obter clareza e engajamento em todos os níveis organizacionais.

Essa abordagem sistemática do gerenciamento de mudanças garante que a implementação da tecnologia seja acompanhada pela maturidade organizacional. Ele cria uma base para confiança, adoção e valor comercial a longo prazo.

Arquitetura para interoperabilidade e colaboração

Implantações de agentes isolados oferecem vitórias locais. Mas o valor corporativo surge quando os agentes podem descobrir, invocar e colaborar uns com os outros de forma dinâmica. Isso significa definir padrões para registro, autenticação e troca de recursos de agentes. Arquitetonicamente, isso

reflete a mudança de monólitos para microsserviços, que são unidades combináveis, reutilizáveis e fracamente acopladas que resolvem problemas complexos em conjunto.

Protocolos emergentes, como A2A e MCP, são fundamentais. Eles permitem a interoperabilidade semântica entre agentes, ferramentas e sistemas de memória. O A2A oferece suporte à interação em nível de pares, o que permite que os agentes negociem a propriedade da tarefa, compartilhem o contexto e coordenem fluxos de trabalho. O MCP complementa isso oferecendo esquemas compartilhados para troca de dados contextuais entre agentes e seus ambientes. Ele padroniza como as funções são invocadas, acessadas e os estados APIs são mantidos. Juntos, esses protocolos promovem extensibilidade, consistência e capacidade de manutenção a longo prazo em todo o ecossistema do agente.

A governança continua sendo crítica. Camadas de controle, como agentes árbitros, permitem a delegação com reconhecimento de políticas sem introduzir gargalos centralizados. Esses agentes atuam como corretores fiduciários. Eles impõem limites enquanto permitem que outros agentes se auto-organizem. A colaboração de agentes ajuda as organizações a escalar seus ecossistemas de IA agentes com agilidade e confiança.

Construindo a governança em um tecido agêntico

Com maior autonomia, surgem maiores riscos. A governança deve ser incorporada à arquitetura do agente desde o primeiro dia. Isso inclui definir limites de políticas que definam o que os agentes podem fazer, aplicar modelos de identidade que determinam de quem eles agem em nome e implementar explicabilidade e rastreabilidade. Os sistemas de observabilidade devem capturar a telemetria do comportamento do agente usando serviços como o Amazon CloudWatch e AWS X-Ray, que fornecem registro centralizado e rastreamento distribuído entre os fluxos de trabalho do agente. Agentes reflexivos podem auditar e avaliar continuamente o desempenho com base nesses feeds de telemetria.

A governança também deve evoluir à medida que o ecossistema do agente amadurece. À medida que os agentes se tornam mais capazes e mais autônomos, os mecanismos de supervisão devem se tornar mais adaptáveis. Atualizações de políticas, controle de capacidade e restrições comportamentais de tempo de execução precisam ser dinâmicas e aplicáveis em grande escala. Confiança não é um recurso incorporado. É continuamente reforçado por meio de arquitetura, comportamento e processo. AWS Identity and Access Management (IAM) e AWS AppConfigdesempenham um papel fundamental na aplicação de identidades seguras, limites de permissão de tempo de execução e alternâncias de comportamento específicas do ambiente entre agentes.

Adotando uma mentalidade operacional que prioriza a tomada de decisões

A automação tradicional se concentra na eficiência do processo, que está executando scripts ou fluxos de trabalho predefinidos com mais rapidez e confiabilidade. A inteligência artificial, por outro lado, introduz a automação que prioriza a tomada de decisões. Os agentes avaliam o contexto, avaliam as opções e adaptam o comportamento em tempo real. Essa mudança de uma mentalidade que prioriza a execução para a tomada de decisão exige um novo pensamento sobre métricas e resultados de sucesso. Em vez de medir o sucesso exclusivamente pela conclusão de tarefas, o sucesso da IA agente é medido pelo quão bem a decisão está alinhada com a intenção, as políticas e as condições em evolução.

Em vez de medir apenas a conclusão da tarefa ou o tempo do ciclo, as organizações devem avaliar a qualidade das decisões e a capacidade de resposta às mudanças. time-to-action KPIs deve incluir métricas como:

- Qualidade da decisão — Até que ponto o agente personalizou sua resposta ao usuário ou cenário específico? Ela tomou decisões diferenciadas que estão alinhadas aos objetivos de negócios e ao contexto do usuário?
- T ime-to-action — Com que rapidez e inteligência o agente avaliou uma situação e respondeu? A latência foi baixa o suficiente para parecer adaptável e humana?
- Descarga cognitiva — quanta análise manual, triagem ou tomada de decisão de rotina o agente foi capaz de realizar em nome de um humano? Isso reduziu o esforço ou apenas o mudou?

As empresas que adotam a mentalidade de tomar decisões em primeiro lugar podem se tornar mais resilientes, adaptáveis e capazes de operar em um novo nível de complexidade.

Dimensionamento com propósito e intenção

Escalar com sucesso a IA de uma agência não significa experimentar mais ferramentas. Trata-se de criar uma camada durável de inteligência corporativa. Isso requer investimentos em infraestrutura de plataforma, cultura operacional, estruturas de governança e alinhamento estratégico. As empresas devem adotar uma abordagem intencional. Eles devem tratar os agentes não como experimentos, mas como componentes centrais de seu modelo operacional digital.

O alinhamento com o AWS Well-Architected Framework ajuda seus sistemas a atender aos padrões corporativos de confiabilidade, segurança, eficiência de desempenho e otimização de custos.

Ferramentas como o SDK do Strands Agents podem acelerar essa jornada fornecendo solicitações estruturadas, registro de ferramentas e prontidão para CI/CD. Isso ajuda as equipes a migrar da experimentação para a entrega escalável usando fluxos de trabalho familiares AWS .

A inteligência artificial não é uma ferramenta; é uma mudança na forma como a inteligência é incorporada às operações. Organizações que se preparam adequadamente podem automatizar mais, operar de forma mais inteligente, se adaptar mais rapidamente e criar vantagens duradouras em um mundo cada vez mais complexo.

Conclusão sobre a operacionalização da IA agente

A IA agente representa mais do que uma mudança tecnológica. Isso marca o surgimento de um novo sistema operacional para a empresa. As organizações que adotam essa transformação vão além dos casos de uso de automação restritos e incorporam inteligência na base de suas operações. Essa mudança trata de redesenhar a forma como as decisões são tomadas, como os sistemas se adaptam e como os resultados são obtidos em grande escala.

Em uma era definida pela crescente complexidade, demanda em tempo real e sobrecarga de informações, o modelo tradicional de automação com script atingiu seus limites. O sucesso agora depende da capacidade de incorporar inteligência diretamente aos fluxos de trabalho para criar sistemas que percebam, raciocinem, ajam e evoluam. A IA agente pode alinhar autonomia com propósito, tomada de decisão com governança e adaptabilidade com responsabilidade.

Essa transição exige uma mudança do pensamento que prioriza a execução para o pensamento que prioriza a decisão. Os sistemas Agentic não seguem simplesmente as instruções. Eles interpretam metas, avaliam compensações e buscam resultados dentro de restrições definidas. Nesse contexto, o sucesso não é medido apenas pela conclusão da tarefa. Também é medido pela qualidade, agilidade e explicabilidade das decisões tomadas em tempo real. As organizações devem repensar as métricas, os incentivos e o design do sistema para apoiar os agentes que operam de forma inteligente sob incertezas.

Operacionalizar a IA agente não é uma atualização. plug-and-play É uma transformação arquitetônica e cultural. Ela exige práticas disciplinadas de gerenciamento do ciclo de vida, fiscalização da confiança, interoperabilidade e alinhamento aos modelos de negócios. Também exige a evolução dos modelos de entrega, como moldar zonas de intenção, incorporar barreiras de tempo de execução e alinhar continuamente o comportamento do agente com os resultados estratégicos. As equipes devem adotar linguagem compartilhada, propriedade compartilhada e responsabilidade compartilhada pelo desempenho e segurança dos agentes.

A prontidão corporativa pode determinar quem prospera nesse novo ambiente. As organizações devem investir em estruturas internas de capacitação, AgentOps capacidades e governança que escalem e criem valor a longo prazo. Aqueles que obtêm sucesso podem criar sistemas mais inteligentes e também podem criar negócios mais adaptáveis, resilientes e orientados por insights.

Este guia estabelece a base. Ele conecta a estratégia à execução e prepara as organizações para criar plataformas escaláveis de agentes inteligentes. A série de conteúdo mais ampla sobre IA agente AWS fornece orientação complementar. Para ver os outros guias desta série, consulte

Agentic AI no site de Orientação AWS Prescritiva. Esta série de conteúdo oferece um roteiro para operacionalizar a autonomia com disciplina e intenção.

Para começar, identifique um espaço de decisão de alto impacto onde os agentes possam oferecer melhorias mensuráveis em velocidade, precisão ou capacidade de resposta. Em seguida, implante um agente piloto focado que tenha ciclos de instrumentação, governança e feedback. Use isso para validar a hipótese de valor, gerar impulso interno e criar confiança na abordagem. O impulso aumenta por meio do aprendizado.

A IA agente não é um destino; é uma camada de capacidade que evolui junto com sua empresa. Isso representa uma mudança de longo prazo em direção à inteligência como infraestrutura. As organizações que lideram nesse espaço podem automatizar mais, responder com mais rapidez, se adaptar melhor e criar modelos operacionais capazes de lidar com a complexidade em escala empresarial.

Recursos para operacionalizar a IA agente

Serviços da AWS

O seguinte Serviços da AWS e os recursos a seguir podem ajudá-lo a criar e operacionalizar sistemas de IA agentes no: Nuvem AWS

- O Amazon API Gateway pode expor os recursos do agente como escaláveis e oferece preços baseados no uso.
- AWS AppConfigoferece gerenciamento de configuração em tempo de execução e alternância de recursos para agentes em todos os locatários ou ambientes.
- O Amazon Bedrock é um serviço de modelo básico que os agentes podem usar para raciocinar, gerar e executar rapidamente.
- AWS Cloud Development Kit (AWS CDK)é uma infraestrutura como serviço de código que você pode usar para implantar e gerenciar pilhas de agentes.
- AWS CloudTrailregistra o histórico de eventos para que você possa acompanhar a atividade do agente, as trilhas de auditoria e os comportamentos de integração.
- A Amazon CloudWatch pode gerenciar registros, métricas e alarmes para monitorar o desempenho do agente e o comportamento de colaboração entre vários agentes.
- AWS CodePipelinefornece CI/CD automação que você pode usar para testar, validar e implantar o código do agente.
- O Amazon Cognito é um serviço de identidade que você pode usar para gerenciar a autenticação de usuários e inquilinos em sistemas multiagentes.
- AWS Configoferece conformidade e detecção de desvios para a política do agente e a configuração do ambiente.
- AWS Cost Explorerpode monitorar o uso em nível de agente e ajudar a alinhar os custos para maximizar seu ROI.
- O Amazon DynamoDB é um serviço de armazenamento que você pode usar para memória de agentes, registros de melhoria e estado contextual.
- O Amazon Elastic File System (Amazon EFS) é um sistema de arquivos compartilhado que você pode usar para colaboração de agentes ou processamento intermediário entre fluxos de trabalho.
- EventBridgeO Amazon é um barramento de eventos principal que você pode usar para rotear tarefas e orquestrar a comunicação na estrutura do agente.

- O Amazon EventBridge Pipes pode agilizar a ingestão e o roteamento de eventos para conectar agentes e serviços.
- A Amazon GuardDuty oferece detecção de ameaças e monitoramento de anomalias que podem apoiar a execução segura de agentes.
- AWS Identity and Access Management (IAM) ajuda você a definir permissões refinadas para execução de agentes e acesso a dados.
- AWS Lambdaé um serviço de computação sem estado que pode executar a lógica de agentes e enxamear drones.
- AWS Marketplaceé uma plataforma de distribuição externa que você pode usar para oferecer recursos de agentes como produtos comerciais.
- AWS Organizationsé um serviço de governança e aplicação de políticas entre contas que pode ajudá-lo a gerenciar a infraestrutura de agentes multilocatários.
- AWS Organizations as políticas de controle de serviços atuam como barreiras para controlar as permissões no nível da conta ou da unidade organizacional.
- O Amazon Quick é uma plataforma generativa de inteligência de negócios (BI) baseada em IA que ajuda você a analisar dados, criar visualizações, automatizar fluxos de trabalho e colaborar com outras pessoas em sua organização.
- AWS Resource Access Manager (AWS RAM) pode ajudá-lo a compartilhar recursos entre contas e serviços de agentes.
- O Amazon SageMaker AI é um serviço que você pode usar para treinamento, ajuste fino e inferência de modelos além dos modelos básicos.
- O Amazon Simple Storage Service (Amazon S3) oferece armazenamento de objetos para bibliotecas imediatas, artefatos de modelos e dados gerados por agentes.
- AWS Step Functionsé um mecanismo de fluxo de trabalho que pode ajudá-lo a coordenar fluxos de multiagentes e pipelines de reciclagem.
- AWS X-Rayoferece rastreamento distribuído que você pode usar para rastrear fluxos de decisão de agentes e dependências de serviços.

Outros AWS recursos

- Fundamentos da IA agêntica em AWS
- Padrões e fluxos de trabalho de IA agentes em AWS
- Estruturas, protocolos e ferramentas de IA da Agentic no AWS

- Construindo arquiteturas sem servidor para IA agente em AWS
- Criação de arquiteturas multilocatárias para IA agente em 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	—	12 de agosto de 2025

AWS Glossário de orientação prescritiva

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

Números

7 Rs

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

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

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

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

A

ABAC

Consulte controle de acesso baseado em atributo.

serviços abstraídos

Veja serviços gerenciados.

ACID

Veja atomicidade, consistência, isolamento, durabilidade.

migração ativa-ativa

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

migração ativa-passiva

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

AGGREGATE FUNCTION

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

AI

Veja inteligência artificial.

AIOps

Veja operações de inteligência artificial.

anonimização

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

antipadrões

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

controle de aplicações

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

portfólio de aplicações

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

inteligência artificial (IA)

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

operações de inteligência artificial (AIOps)

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

criptografia assimétrica

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

atomicidade, consistência, isolamento, durabilidade (ACID)

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

controle de acesso por atributo (ABAC)

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

fonte de dados autorizada

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

Zona de disponibilidade

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

AWS Estrutura de adoção da nuvem (AWS CAF)

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

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

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

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

B

bot malicioso

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

BCP

Veja planejamento de continuidade de negócios

gráfico de comportamento

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

sistema big-endian

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

classificação binária

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

filtro de bloom

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

blue/green deployment (implantação azul/verde)

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

bot

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

botnet

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

ramo

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

Acesso de emergência

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

estratégia brownfield

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

cache do buffer

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

capacidade de negócios

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

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

planejamento de continuidade de negócios (BCP)

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

C

CAF

Veja AWS Cloud Adoption Framework.

implantação canário

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

CCoE

Veja Centro de Excelência da Nuvem.

CDC

Veja captura de dados de alteração.

captura de dados de alterações (CDC)

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

engenharia do caos

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

CI/CD

Veja integração e entrega contínuas.

classificação

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

criptografia no lado do cliente

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

Centro de excelência em nuvem (CCoE)

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

computação em nuvem

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

modelo operacional em nuvem

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

estágios de adoção da nuvem

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

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

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

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

CMDB

Veja banco de dados de gerenciamento de configuração.

repositório de código

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

cache frio

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

dados frios

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

visão computacional (CV)

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

desvio de configuração

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

banco de dados de gerenciamento de configuração (CMDB)

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

pacote de conformidade

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

integração contínua e entrega contínua (CI/CD)

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

CV

Veja visão computacional.

D

dados em repouso

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

classificação de dados

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

desvio de dados

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

dados em trânsito

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

data mesh

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

minimização de dados

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

perímetro de dados

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

pré-processamento de dados

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

proveniência dos dados

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

titular dos dados

Um indivíduo cujos dados estão sendo coletados e processados.

data warehouse

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

linguagem de definição de dados (DDL)

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

linguagem de manipulação de dados (DML)

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

DDL

Veja linguagem de definição de banco de dados.

deep ensemble

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

Aprendizado profundo

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

defense-in-depth

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

administrador delegado

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

implantação

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

ambiente de desenvolvimento

Veja ambiente.

controle detectivo

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

mapeamento do fluxo de valor de desenvolvimento (DVSM)

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

gêmeo digital

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

tabela de dimensões

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

desastre

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

Recuperação de desastres (RD)

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

DML

Veja linguagem de manipulação de banco de dados.

design orientado por domínio

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

DR

Veja recuperação de desastres.

Detecção da oscilação

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

DVSM

Veja mapeamento do fluxo de valor de desenvolvimento.

E

EDA

Veja análise exploratória de dados.

EDI

Veja intercâmbio eletrônico de dados.

computação de borda

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

intercâmbio eletrônico de dados (EDI)

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

criptografia

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

chave de criptografia

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

endianismo

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

endpoint

Veja endpoint de serviço.

serviço de endpoint

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

planejamento de recursos empresariais (ERP)

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

criptografia envelopada

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

ambiente

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

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

epic

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

ERP

Veja planejamento de recursos empresariais.

análise exploratória de dados (EDA)

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

F

tabela de fatos

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

Antecipar-se à falha

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

delimitação de isolamento contra falhas

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

ramificação de recursos

Veja ramificação.

recursos

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

importância do recurso

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

transformação de recursos

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

prompt few shot

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

FGAC

Veja controle de acesso refinado.

Controle de acesso refinado (FGAC)

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

migração flash-cut

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

FM

Veja modelo de base.

modelo de base (FM)

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

G

IA generativa

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

bloqueio geográfico

Veja restrições geográficas.

restrições geográficas (bloqueio geográfico)

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

Fluxo de trabalho do GitFlow

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

golden image

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

estratégia greenfield

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

barreira de proteção

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

H

HA

Veja alta disponibilidade.

migração heterogênea de bancos de dados

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

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

alta disponibilidade (HA)

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

modernização de historiador

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

dados de hold-out

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

migração homogênea de bancos de dados

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

dados quentes

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

hotfix

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

período de hipercuidados

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

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

eu

IaC

Veja infraestrutura como código.

Política baseada em identidade

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

aplicação ociosa

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

IIoT

Veja Internet das Coisas Industrial.

infraestrutura imutável

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

VPC de entrada (admissão)

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

migração incremental

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

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

Indústria 4.0

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

infraestrutura

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

Infraestrutura como código (IaC)

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

Internet industrial das coisas (IIoT)

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

VPC de inspeção

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

Internet das coisas (IoT)

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

interpretabilidade

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

IoT

Veja Internet das Coisas.

Biblioteca de informações de TI (ITIL)

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

Gerenciamento de serviços de TI (ITSM)

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

ITIL

Veja biblioteca de informações de TI.

ITSM

Veja gerenciamento de serviços de TI.

L

controle de acesso baseado em etiqueta (LBAC)

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

zona de pouso

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

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

grande modelo de linguagem (LLM)

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

migração de grande porte

Uma migração de 300 servidores ou mais.

LBAC

Veja controle de acesso baseado em rótulo.

privilégio mínimo

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

mover sem alterações (lift-and-shift)

Veja 7 Rs.

sistema little-endian

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

LLM

Veja grande modelo de linguagem.

ambientes inferiores

Veja ambiente.

M

machine learning (ML)

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

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

ramificação principal

Veja ramificação.

Malware

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

Serviços gerenciados

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

sistema de execução de manufatura (MES)

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

MAP

Veja Programa de Aceleração da Migração.

mecanismo

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

conta de membro

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

MES

Veja sistema de execução de manufatura.

Transporte de Telemetria de Enfileiramento de Mensagens (MQTT)

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

microsserviço

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

arquitetura de microsserviços

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

Programa de Aceleração da Migração (MAP)

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

migração em escala

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

fábrica de migração

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

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

metadados de migração

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

padrão de migração

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

Avaliação de Portfólio para Migração (MPA)

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

Avaliação de Preparação para Migração (MRA)

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

estratégia de migração

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

ML

Veja machine learning.

modernização

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

avaliação de preparação para modernização

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

aplicações monolíticas (monólitos)

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

MPA

Veja Avaliação do Portfólio para Migração.

MQTT

Veja Transporte de Telemetria de Enfileiramento de Mensagens.

classificação multiclasse

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

infraestrutura mutável

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

O

OAC

Veja controle de acesso de origem.

OAI

Veja identidade de acesso de origem.

OCM

Veja gerenciamento de alterações organizacionais.

migração offline

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

OI

Veja integração de operações.

Ola

Veja acordo de nível operacional.

migração online

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

OPC-UA

Veja Open Process Communications - Unified Architecture.

Open Process Communications - Unified Architecture (OPC-UA)

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

acordo de nível operacional (OLA)

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

análise de prontidão operacional (ORR)

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

tecnologia operacional (TO)

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

integração de operações (OI)

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

trilha organizacional

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

gerenciamento de alterações organizacionais (OCM)

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

controle de acesso de origem (OAC)

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

Identidade do acesso de origem (OAI)

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

ORR

Veja análise de prontidão operacional.

OT

Veja tecnologia operacional.

VPC de saída (egresso)

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

P

limite de permissões

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

Informações de identificação pessoal (PII)

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

PII

Veja informações de identificação pessoal.

manual

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

PLC

Veja controlador lógico programável.

PLM

Veja gerenciamento do ciclo de vida do produto.

política

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

persistência poliglota

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

avaliação do portfólio

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

predicado

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

pushdown de predicados

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

controle preventivo

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

principal (entidade principal)

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

Privacidade por design

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

zonas hospedadas privadas

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

controle proativo

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

gerenciamento do ciclo de vida do produto (PLM)

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

ambiente de produção

Veja ambiente.

controlador lógico programável (PLC)

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

encadeamento de prompts

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

pseudonimização

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

publish/subscribe (pub/sub)

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

Q

plano de consulta

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

regressão de planos de consultas

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

R

Matriz RACI

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

RAG

Veja geração aumentada via recuperação.

ransomware

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

Matriz RASCI

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

RCAC

Veja controle de acesso por linha e coluna.

réplica de leitura

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

Redefinir arquitetura

Veja 7 Rs.

objetivo de ponto de recuperação (RPO).

O máximo período de tempo aceitável desde o último ponto de recuperação de dados. Isso determina o que é considerado uma perda aceitável de dados entre o último ponto de recuperação e a interrupção do serviço.

objetivo de tempo de recuperação (RTO)

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

refatorar

Veja 7 Rs.

Região

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

regressão

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

redefinir a hospedagem

Veja 7 Rs.

versão

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

realocar

Veja 7 Rs.

redefinir a plataforma

Veja 7 Rs.

recomprar

Veja 7 Rs.

resiliência

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

política baseada em recurso

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

matriz responsável, accountable, consultada, informada (RACI)

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

controle responsivo

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

reter

Veja 7 Rs.

Retirada

Veja 7 Rs.

Geração Aumentada de Recuperação (RAG)

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

alternância

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

controle de acesso por linha e coluna (RCAC)

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

RPO

Veja objetivo de ponto de recuperação.

RTO

Veja objetivo de tempo de recuperação.

runbook

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

S

SAML 2.0

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

SCADA

Veja controle de supervisão e aquisição de dados.

SCP

Veja política de controle de serviço.

secret

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

segurança desde a concepção

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

controle de segurança

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

hardening da segurança

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

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

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

automação de resposta de segurança

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

Criptografia do lado do servidor

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

política de controle de serviços (SCP)

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

service endpoint (endpoint de serviço)

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

acordo de serviço (SLA)

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

indicador de nível de serviço (SLI)

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

objetivo de nível de serviço (SLO)

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

modelo de responsabilidade compartilhada

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

SIEM

Veja sistema de gerenciamento de eventos e informações de segurança.

ponto único de falha (SPOF)

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

SLA

Veja cacordo de serviço.

SLI

Veja indicador de nível de serviço.

SLO

Veja objetivo de nível de serviço.

split-and-seed modelo

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

SPOF

Veja ponto único de falha.

esquema em estrela

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

padrão strangler fig

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

sub-rede

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

controle supervisório e aquisição de dados (SCADA)

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

symmetric encryption (criptografia simétrica)

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

testes sintéticos

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

prompt do sistema

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

T

tags

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

variável-alvo

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

lista de tarefas

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

ambiente de teste

Veja ambiente.

treinamento

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

gateway de trânsito

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

fluxo de trabalho baseado em troncos

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

Acesso confiável

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

tuning (ajustar)

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

equipe de duas pizzas

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

U

incerteza

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

tarefas indiferenciadas

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

ambientes superiores

Veja ambiente.

V

aspiração

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

controle de versões

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

emparelhamento da VPC

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

Vulnerabilidade

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

W

cache quente

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

dados mornos

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

função de janela

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

workload

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

workstreams

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

WORM

Veja gravação única e várias leituras.

WQF

Veja AWS Workload Qualification Framework.

gravação única e várias leituras (WORM)

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

Z

exploração de dia zero

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

vulnerabilidade de dia zero

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

prompt zero shot

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

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

aplicação zumbi

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

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