Estruturas, plataformas, protocolos e ferramentas de IA da Agentic no AWS

AWS Orientação prescritiva

AWS Orientação prescritiva: Estruturas, plataformas, protocolos e ferramentas de IA da Agentic no AWS

Table of Contents

Estruturas, plataformas, protocolos e ferramentas de IA da Agentic no AWS

Aaron Sempf, Ansley Verzosa e Joshua Samuel, da Amazon Web Services ()AWS

Janeiro de 2026 (histórico do documento)

A IA agente é um paradigma poderoso na interseção entre IA, sistemas distribuídos e engenharia de software. É uma classe de sistemas inteligentes composta por agentes de software autônomos e assíncronos que usam modelos de IA e se integram a ferramentas e recursos. Os agentes demonstram agência, podem perceber o contexto, raciocinar sobre metas, tomar decisões e realizar ações propositais em nome de usuários ou sistemas. Esses agentes operam de forma independente, geralmente de forma colaborativa, em ambientes distribuídos e são projetados para perseguir objetivos delegados com inteligência, memória e intenção incorporadas.

Além AWS disso, as organizações podem aproveitar a IA agente para automatizar fluxos de trabalho complexos, aprimorar os processos de tomada de decisão e criar sistemas mais responsivos. Este guia fornece informações sobre os principais componentes necessários para criar soluções eficazes de IA agêntica:

- O Frameworks traça o perfil das estruturas atuais de IA para agentes, incluindo análises de seus benefícios e casos de uso. Saiba como essas estruturas reduzem o trabalho pesado indiferenciado entre padrões, protocolos e ferramentas. Entenda os principais critérios de seleção para escolher a estrutura certa para suas necessidades.
- As plataformas fornecem uma visão geral das plataformas de IA agentes (agente gerenciado, orquestração de código aberto e híbrida) e considerações para seleção ou design.
- Protocolos explora protocolos de comunicação padronizados essenciais para interações de agentes. Agent-to-agentprotocolos estão surgindo, como o Model Context Protocol (MCP) e o Agent2Agent (A2A) de código aberto, junto com outras implementações proprietárias. Descubra como protocolos comuns permitem que diferentes protocolos interajam perfeitamente.
- As ferramentas fornecem informações sobre ferramentas baseadas em protocolos (como o MCP), ferramentas nativas da estrutura e meta-ferramentas. As organizações podem criar um kit de ferramentas que se integre aos principais sistemas em seus fluxos de trabalho, permitindo fluxos de trabalho agentes baseados no usuário final e no servidor.

Público-alvo

Este guia é para arquitetos, desenvolvedores e líderes de tecnologia que buscam aproveitar o poder dos agentes de software orientados por IA em aplicativos modernos nativos da nuvem.

Objetivos

Este guia ajuda você a:

- Compare diferentes estruturas de IA agente para selecionar a mais adequada para seu caso de uso.
- Saiba mais sobre plataformas de IA agênticas que fornecem recursos para transformar agentes individuais em sistemas coordenados e adaptáveis.
- Entenda as vantagens dos protocolos abertos para criar arquiteturas sustentáveis de IA agente.
- Crie uma estratégia apropriada de integração de ferramentas ao criar sistemas de agentes.

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.

Estruturas

O Foundations of Agentic AI on AWS examina os principais padrões e fluxos de trabalho que permitem um comportamento autônomo e direcionado a objetivos. No centro da implementação desses padrões está a escolha da estrutura. Uma estrutura é a base de software do código pré-escrito que fornece um ambiente estruturado e funcionalidade comum para criar e gerenciar as ferramentas e os recursos de orquestração necessários para criar agentes de IA autônomos prontos para produção.

Estruturas de IA agênticas eficazes fornecem vários recursos essenciais que transformam as interações brutas do modelo de linguagem grande (LLM) em sistemas coordenados e inteligentes capazes de raciocinar, colaborar e agir:

- A orquestração de agentes coordena o fluxo de informações e a tomada de decisões em um ou vários agentes para atingir metas complexas sem intervenção humana.
- A integração de ferramentas permite que os agentes interajam com sistemas externos e fontes de dados para ampliar seus recursos além do processamento de linguagem. APIs Para obter mais informações, consulte Visão geral das ferramentas na Strands Agents documentação.
- O gerenciamento de memória fornece um estado persistente ou baseado em sessão para manter o contexto em todas as interações, essencial para tarefas adaptáveis ou de longa duração. Estruturas mais avançadas incorporam memória de longo prazo para armazenar resumos e preferências do usuário, permitindo experiências agentes personalizadas e contextualmente conscientes. Para obter mais informações, consulte Como pensar sobre estruturas de agentes no LangChain blog.
- A definição do fluxo de trabalho suporta padrões estruturados, como cadeias, roteamento, paralelização e ciclos de reflexão, que permitem um raciocínio autônomo sofisticado.
- A implantação e o monitoramento facilitam a transição do desenvolvimento para a produção com observabilidade para sistemas autônomos. Para obter mais informações, consulte o anúncio de disponibilidade AgentCore geral do Amazon Bedrock.

Esses recursos são implementados com abordagens e ênfases variadas em todo o cenário da estrutura, cada uma oferecendo vantagens distintas para diferentes casos de uso de agentes autônomos e contextos organizacionais.

Esta seção traça o perfil e compara as principais estruturas para criar soluções de IA agênticas, com foco em seus pontos fortes, limitações e casos de uso ideais para operação autônoma:

- Agentes de filamentos
- LangChain and LangGraph
- Tripulação AI
- AutoGen
- ???
- Comparando estruturas de IA agênticas

> **Note**
>
> Esta seção aborda as estruturas que apoiam especificamente a agência da IA e não abrange interfaces de front-end ou IA generativa sem agência.

Strands Agents

Strands Agentsé um SDK de código aberto que foi lançado inicialmente pela AWS, conforme descrito no AWS Open Source Blog. Strands Agentsfoi projetado para criar agentes de IA autônomos com uma abordagem que prioriza o modelo. Ele fornece uma estrutura flexível e extensível, projetada para funcionar perfeitamente e, ao Serviços da AWS mesmo tempo, permanecer aberta à integração com componentes de terceiros. O Strands Agents é ideal para criar soluções totalmente autônomas.

Principais características do Strands Agents

Strands Agentsinclui os seguintes recursos principais:

- Design que prioriza o modelo — Construído com base no conceito de que o modelo básico é o núcleo da inteligência do agente, permitindo um raciocínio autônomo sofisticado. Para obter mais informações, consulte Agent Loop na Strands Agents documentação.
- Padrões de colaboração multiagente — Modelos de coordenação integrados, como padrões Swarm, Graph e Workflow, que permitem colaboração e governança escaláveis em redes de agentes distribuídos. Para obter mais informações, consulte Padrões de vários agentes na documentação do Strands Agents.
- Integração MCP — Suporte nativo para o Model Context Protocol (MCP), permitindo o fornecimento de contexto padronizado LLMs para uma operação autônoma consistente.

- AWS service (Serviço da AWS) integração — conexão perfeita com Amazon Bedrock,, AWS Lambda AWS Step Functions, e outros Serviços da AWS para fluxos de trabalho autônomos abrangentes. Para obter mais informações, consulte Resumo AWS semanal (AWS blog).
- Seleção de modelos básicos — Suporta vários modelos básicos, incluindo Anthropic Claude, Amazon Nova (Premier, Pro, Lite e Micro) no Amazon Bedrock e outros, para otimizar diferentes capacidades de raciocínio autônomo. Para obter mais informações, consulte Amazon Bedrock na Strands Agents documentação.
- Integração da API LLM — Integração flexível com diferentes interfaces de serviço LLM, incluindo Amazon Bedrock, OpenAI e outras, para implantação em produção. Para obter mais informações, consulte Uso básico do Amazon Bedrock na Strands Agents documentação.
- Capacidades multimodais — Support para várias modalidades, incluindo processamento de texto, fala e imagem para interações abrangentes com agentes autônomos. Para obter mais informações, consulte Amazon Bedrock Multimodal Support na Strands Agents documentação.
- Ecossistema de ferramentas — Conjunto rico de ferramentas para AWS service (Serviço da AWS) interação, com extensibilidade para ferramentas personalizadas que expandem as capacidades autônomas. Para obter mais informações, consulte Visão geral das ferramentas na Strands Agents documentação.

Quando utilizar Strands Agents

Strands Agentsé particularmente adequado para cenários de agentes autônomos, incluindo:

- Organizações que se baseiam em uma AWS infraestrutura que desejam integração nativa com fluxos Serviços da AWS de trabalho autônomos
- Equipes que exigem recursos de segurança, escalabilidade e conformidade de nível empresarial para sistemas autônomos de produção
- Projetos que precisam de flexibilidade na seleção de modelos em diferentes fornecedores para tarefas autônomas especializadas
- Casos de uso que exigem forte integração com AWS fluxos de trabalho e recursos existentes para processos autônomos de ponta a ponta

Abordagem de implementação para Strands Agents

Strands Agentsfornece uma abordagem de implementação direta para as partes interessadas da empresa, conforme descrito em seu Guia de Início Rápido. A estrutura permite que as organizações:

- Selecione modelos básicos como o Amazon Nova (Premier, Pro, Lite ou Micro) no Amazon Bedrock com base em requisitos comerciais específicos.
- Defina ferramentas personalizadas que se conectem a sistemas corporativos e fontes de dados.
- Processe várias modalidades, incluindo texto, imagens e fala.
- Implante agentes que possam responder de forma autônoma às consultas comerciais e realizar tarefas.

Essa abordagem de implementação permite que as equipes de negócios desenvolvam e implantem rapidamente agentes autônomos sem profundo conhecimento técnico no desenvolvimento de modelos de IA.

Exemplo real de Strands Agents

AWS Transform for .NET usa Strands Agents para potencializar seus recursos de modernização de aplicativos, conforme descrito em AWS Transform for .NET, o primeiro serviço de IA agente para modernizar aplicativos.NET em grande escala (AWS Blog). Este serviço de produção emprega vários agentes autônomos especializados. Os agentes trabalham juntos para analisar aplicativos.NET legados, planejar estratégias de modernização e executar transformações de código em arquiteturas nativas da nuvem sem intervenção humana. AWS Transform for .NET demonstra a prontidão de produção de Strands Agents sistemas autônomos corporativos.

LangChain e LangGraph

LangChainé uma das estruturas mais estabelecidas no ecossistema de IA agente. LangGraphamplia seus recursos para oferecer suporte a fluxos de trabalho de agentes complexos e dinâmicos, conforme descrito no LangChain Blog. Juntos, eles fornecem uma solução abrangente para criar agentes de IA autônomos sofisticados com recursos avançados de orquestração para operação independente.

Principais características de LangChain e LangGraph

LangChaine LangGraph inclua os seguintes recursos principais:

- Ecossistema de componentes — Ampla biblioteca de componentes pré-construídos para vários recursos de agentes autônomos, permitindo o rápido desenvolvimento de agentes especializados. Para obter mais informações, consulte Início rápido na LangChain documentação.

- Seleção de modelos básicos — Suporte para diversos modelos de fundação, incluindo modelos Anthropic Claude, Amazon Nova (Premier, Pro, Lite e Micro) no Amazon Bedrock e outros para diferentes capacidades de raciocínio. Para obter mais informações, consulte Entradas e saídas na LangChain documentação.
- Integração da API LLM — Interfaces padronizadas para vários provedores de serviços de modelo de linguagem grande (LLM), incluindo Amazon Bedrock e outrosOpenAI, para implantação flexível. Para obter mais informações, consulte a LLMs documentação do LangChain.
- Processamento multimodal — Suporte integrado para processamento de texto, imagem e áudio para permitir interações ricas com agentes autônomos multimodais. Para obter mais informações, consulte Multimodalidade na LangChain documentação.
- Fluxos de trabalho baseados em gráficos — LangGraph permitem definir comportamentos complexos de agentes autônomos como máquinas de estado, suportando uma lógica de decisão sofisticada. Para obter mais informações, consulte o anúncio da LangGraphPlatform GA.
- Abstrações de memória — Várias opções para gerenciamento de memória de curto e longo prazo, o que é essencial para agentes autônomos que mantêm o contexto ao longo do tempo. Para obter mais informações, consulte Como adicionar memória aos chatbots na LangChain documentação.
- Integração de ferramentas — Ecossistema rico de integrações de ferramentas em vários serviços e ampliando APIs as capacidades de agentes autônomos. Para obter mais informações, consulte Ferramentas na LangChain documentação.
- LangGraph plataforma — Solução gerenciada de implantação e monitoramento para ambientes de produção, oferecendo suporte a agentes autônomos de longa duração. Para obter mais informações, consulte o anúncio da LangGraphPlatform GA.

Quando usar LangChain e LangGraph

LangChaine LangGraph são particularmente adequados para cenários de agentes autônomos, incluindo:

- Fluxos de trabalho complexos de raciocínio em várias etapas que exigem orquestração sofisticada para tomada de decisão autônoma
- Projetos que precisam de acesso a um grande ecossistema de componentes e integrações pré-construídos para diversas capacidades autônomas
- Equipes com infraestrutura e experiência Python em aprendizado de máquina (ML) existentes que desejam criar sistemas autônomos

- Casos de uso que exigem gerenciamento complexo de estados em sessões de agentes autônomos de longa duração

Abordagem de implementação para LangChain e LangGraph

LangChaine LangGraph forneça uma abordagem de implementação estruturada para as partes interessadas da empresa, conforme detalhado na LangGraphdocumentação. A estrutura permite que as organizações:

- Defina gráficos de fluxo de trabalho sofisticados que representem os processos de negócios.
- Crie padrões de raciocínio em várias etapas com pontos de decisão e lógica condicional.
- Integre recursos de processamento multimodal para lidar com diversos tipos de dados.
- Implemente o controle de qualidade por meio de mecanismos integrados de revisão e validação.

Essa abordagem baseada em gráficos permite que as equipes de negócios modelem processos de decisão complexos como fluxos de trabalho autônomos. As equipes têm uma visibilidade clara de cada etapa do processo de raciocínio e a capacidade de auditar os caminhos de decisão.

Exemplo real de e LangChain LangGraph

Vodafoneimplementou agentes autônomos usando LangChain (eLangGraph) para aprimorar seus fluxos de trabalho de engenharia de dados e operações, conforme detalhado em seu estudo de caso LangChain corporativo. Eles criaram assistentes internos de IA que monitoram de forma autônoma as métricas de desempenho, recuperam informações dos sistemas de documentação e apresentam insights acionáveis, tudo por meio de interações em linguagem natural.

A Vodafone implementação usa carregadores LangChain modulares de documentos, integração vetorial e suporte para vários LLMs (OpenAI, LLaMA 3 eGemini) para prototipar e comparar rapidamente esses pipelines. Em seguida, LangGraph costumavam estruturar a orquestração de vários agentes implantando subagentes modulares. Esses agentes realizam tarefas de coleta, processamento, resumo e raciocínio. LangGraphintegrou esses agentes APIs em seus sistemas em nuvem.

CrewAI

CrewAIé uma estrutura de código aberto focada especificamente na orquestração autônoma de vários agentes, disponível em. GitHub Ele fornece uma abordagem estruturada para criar equipes de

agentes autônomos especializados que colaboram para resolver tarefas complexas sem intervenção humana. CrewAIenfatiza a coordenação baseada em funções e a delegação de tarefas.

Principais características do CrewAI

CrewAIfornece os seguintes recursos principais:

- Design de agente baseado em funções — Agentes autônomos são definidos com funções, metas e histórias de fundo específicas para permitir conhecimentos especializados. Para obter mais informações, consulte Criação de agentes eficazes na CrewAI documentação.
- Delegação de tarefas — Mecanismos integrados para atribuir tarefas de forma autônoma aos agentes apropriados com base em suas capacidades. Para obter mais informações, consulte Tarefas na CrewAI documentação.
- Colaboração de agentes — Estrutura para comunicação autônoma entre agentes e compartilhamento de conhecimento sem mediação humana. Para obter mais informações, consulte Colaboração na CrewAI documentação.
- Gerenciamento de processos — fluxos de trabalho estruturados para execução sequencial e paralela de tarefas autônomas. Para obter mais informações, consulte Processos na CrewAI documentação.
- Seleção de modelos básicos — Suporte para vários modelos básicos, incluindo modelos Anthropic Claude, Amazon Nova (Premier, Pro, Lite e Micro) no Amazon Bedrock e outros para otimizar diferentes tarefas de raciocínio autônomo. Para obter mais informações, consulte a LLMs documentação do CrewAI.
- Integração da API LLM — Integração flexível com várias interfaces de serviço LLM, OpenAI incluindo Amazon Bedrock e implantações de modelos locais. Para obter mais informações, consulte Exemplos de configuração do provedor na CrewAI documentação.
- Suporte multimodal — Capacidades emergentes para lidar com texto, imagem e outras modalidades para interações abrangentes de agentes autônomos. Para obter mais informações, consulte Usando agentes multimodais na CrewAI documentação.

Quando utilizar CrewAI

CrewAIé particularmente adequado para cenários de agentes autônomos, incluindo:

- Problemas complexos que se beneficiam da experiência especializada e baseada em funções trabalhando de forma autônoma

- Projetos que exigem colaboração explícita entre vários agentes autônomos
- Casos de uso em que a decomposição de problemas baseada em equipe melhora a resolução autônoma de problemas
- Cenários que exigem uma separação clara de preocupações entre diferentes funções de agentes autônomos

Abordagem de implementação para CrewAI

CrewAIfornece uma implementação baseada em funções da abordagem de equipes de agentes de IA para as partes interessadas da empresa, conforme detalhado em Introdução na CrewAI documentação. A estrutura permite que as organizações:

- Defina agentes autônomos especializados com funções, metas e áreas de especialização específicas.
- Atribua tarefas aos agentes com base em suas capacidades especializadas.
- Estabeleça dependências claras entre as tarefas para criar fluxos de trabalho estruturados.
- Organize a colaboração entre vários agentes para resolver problemas complexos.

Essa abordagem baseada em funções reflete as estruturas de equipes humanas, tornando-a intuitiva para os líderes de negócios entenderem e implementarem. As organizações podem criar equipes autônomas com áreas de especialização especializadas que colaboram para alcançar os objetivos de negócios, da mesma forma que as equipes humanas operam. No entanto, a equipe autônoma pode trabalhar continuamente sem intervenção humana.

Exemplo real de CrewAI

AWS implementou sistemas multiagentes autônomos usando o CrewAI integrado ao Amazon Bedrock, conforme detalhado no estudo de caso publicado. CrewAI AWS e CrewAI desenvolveu uma estrutura segura e neutra em relação ao fornecedor. A arquitetura “CrewAIflows-and-crews” de código aberto se integra perfeitamente aos modelos básicos, sistemas de memória e barreiras de conformidade do Amazon Bedrock.

Os principais elementos da implementação incluem:

- Planos e código aberto — AWS e designs de referência CrewAI lançados que mapeiam CrewAI agentes para modelos e ferramentas de observabilidade do Amazon Bedrock. Eles também

lançaram sistemas exemplares, como uma equipe de auditoria de AWS segurança multiagente, fluxos de modernização de código e automação de back-office de bens de consumo embalados (CPG).

- Integração da pilha de observabilidade — A solução incorpora monitoramento com a Amazon e CloudWatch, AgentOps permitindo a rastreabilidade e a LangFuse depuração, desde a prova de conceito até a produção.
- Retorno sobre o investimento (ROI) demonstrado — Os primeiros pilotos mostram grandes melhorias — execução 70% mais rápida para um grande projeto de modernização de código e cerca de 90% de redução no tempo de processamento de um fluxo de back-office de CPG.

AutoGen

AutoGené uma estrutura de código aberto que foi lançada inicialmente pelaMicrosoft. AutoGenconcentra-se em capacitar agentes de IA autônomos conversacionais e colaborativos. Ele fornece uma arquitetura flexível para criar sistemas multiagentes com ênfase em interações assíncronas e orientadas por eventos entre agentes para fluxos de trabalho autônomos complexos.

Principais características do AutoGen

AutoGenfornece os seguintes recursos principais:

- Agentes conversacionais — Construídos em torno de conversas em linguagem natural entre agentes autônomos, permitindo um raciocínio sofisticado por meio do diálogo. Para obter mais informações, consulte Estrutura de conversação multiagente na AutoGen documentação.
- Arquitetura assíncrona — design orientado por eventos para interações de agentes autônomos sem bloqueio, suportando fluxos de trabalho paralelos complexos. Para obter mais informações, consulte Resolvendo várias tarefas em uma sequência de bate-papos assíncronos na documentação. AutoGen
- H uman-in-the-loop — Forte suporte à participação humana opcional em fluxos de trabalho de agentes autônomos, quando necessário. Para obter mais informações, consulte Permitir feedback humano em agentes na AutoGen documentação.
- Geração e execução de código — Recursos especializados para agentes autônomos focados em código que podem escrever e executar código. Para obter mais informações, consulte Execução de código na AutoGen documentação.

- Comportamentos personalizáveis — Configuração flexível de agentes autônomos e controle de conversação para diversos casos de uso. Para obter mais informações, consulte agentchat.conversable_agent na documentação. AutoGen
- Seleção de modelos básicos — Suporte para vários modelos básicos, incluindo modelos Anthropic Claude, Amazon Nova (Premier, Pro, Lite e Micro) no Amazon Bedrock e outros para diferentes capacidades de raciocínio autônomo. Para obter mais informações, consulte Configuração do LLM na AutoGen documentação.
- Integração da API LLM — Configuração padronizada para várias interfaces de serviço LLM, incluindo Amazon Bedrock, e. OpenAI Azure OpenAI Para obter mais informações, consulte oai.openai_utils na Referência da API. AutoGen
- Processamento multimodal — Support para processamento de texto e imagem para permitir interações ricas com agentes autônomos multimodais. Para obter mais informações, consulte Envolvendo-se com modelos multimodais: GPT-4V na documentação. AutoGen AutoGen

Quando utilizar AutoGen

AutoGené particularmente adequado para cenários de agentes autônomos, incluindo:

- Aplicativos que exigem fluxos conversacionais naturais entre agentes autônomos para raciocínio complexo
- Projetos que precisam de operação totalmente autônoma e recursos opcionais de supervisão humana
- Casos de uso que envolvem geração, execução e depuração autônomas de código sem intervenção humana
- Cenários que exigem padrões de comunicação de agentes autônomos flexíveis e assíncronos

Abordagem de implementação para AutoGen

AutoGenfornece uma abordagem de implementação conversacional para as partes interessadas da empresa, conforme detalhado em Introdução na AutoGen documentação. A estrutura permite que as organizações:

- Crie agentes autônomos que se comunicam por meio de conversas em linguagem natural.
- Implemente interações assíncronas e orientadas por eventos entre vários agentes.

- Combine operação totalmente autônoma com supervisão humana opcional quando necessário.
- Desenvolva agentes especializados para diferentes funções de negócios que colaborem por meio do diálogo.

Essa abordagem conversacional torna o raciocínio do sistema autônomo transparente e acessível aos usuários corporativos. Os tomadores de decisão podem observar o diálogo entre os agentes para entender como as conclusões são alcançadas e, opcionalmente, participar da conversa quando o julgamento humano é necessário.

Exemplo real de AutoGen

Magentic-Oneé um sistema multiagente generalista e de código aberto projetado para resolver de forma autônoma tarefas complexas de várias etapas em diversos ambientes, conforme descrito no blog AI Frontiers. Microsoft Em sua essência, está o agente Orchestrator, que decompõe metas de alto nível e acompanha o progresso usando livros contábeis estruturados. Esse agente delega subtarefas a agentes especializados (comoWebSurfer, FileSurferCoder, eComputerTerminal) e se adapta dinamicamente replanejando quando necessário.

O sistema é baseado na AutoGen estrutura e é independente do modelo, usando como padrão o GPT-4o. Ele alcança desempenho de última geração em benchmarks como, e —tudo sem ajustes específicos da tarefa. GAIA AssistantBench WebArena Além disso, ele oferece suporte à extensibilidade modular e à avaliação rigorosa por meio de sugestões. AutoGenBench

LlamaIndex

LlamaIndexé uma estrutura de dados projetada especificamente para conectar grandes modelos de linguagem (LLMs) a fontes de dados externas para permitir aplicativos sofisticados de geração aumentada de recuperação (RAG) e inteligência artificial. A estrutura fornece abstrações e fluxos de trabalho de desenvolvimento acelerados para sistemas agentes, padrões de orquestração personalizados e integrações de sistemas que reduzem as soluções de IA orientadas pelo conhecimento. time-to-production

Principais características do LlamaIndex

LlamaIndexfornece um conjunto abrangente de recursos que o torna particularmente adequado para aplicativos de IA de agentes corporativos:

- Arquitetura centrada em dados — se destaca na ingestão, indexação e recuperação de informações de mais de 100 formatos de dados, incluindo documentos do Microsoft Word PDFs, planilhas e muito mais. A estrutura transforma dados corporativos em bases de conhecimento consultáveis que são otimizadas para agentes de IA. Para obter mais informações, consulte a documentação do LlamaIndex.
- Implantação pronta para produção — LlamaIndex oferece estruturas de código aberto e serviços gerenciados por meio de recursos de nível corporativoLlamaCloud, incluindo controles de segurança, escalabilidade, integrações de observabilidade e flexibilidade de implantação. Para obter mais informações, consulte a documentação da LlamaIndex estrutura.
- Processamento avançado de documentos — LlamaCloud fornece recursos de análise, extração, indexação e recuperação de documentos que lidam com layouts complexos, tabelas aninhadas, conteúdo multimodal e até anotações manuscritas. Essa análise sofisticada permite que os agentes trabalhem de forma eficaz com documentos corporativos reais que contêm gráficos, diagramas e formatação complexa. Para obter mais informações, consulte a documentação do LlamaCloud.
- Orquestração de fluxos de trabalho — LlamaAgents fornece um mecanismo de orquestração assíncrono e orientado por eventos para criar sistemas agentes de várias etapas. Os fluxos de trabalho oferecem suporte a padrões complexos, incluindo loops, execução paralela, ramificação condicional e retomada com estado, o que os torna ideais para interações sofisticadas com agentes. Para obter mais informações, consulte a documentação dos LlamaIndex fluxos de trabalho.
- Capacidades de recuperação agente — modos de recuperação avançados, incluindo pesquisa híbrida, pesquisa semântica e roteamento automático que determinam de forma inteligente a melhor estratégia de recuperação para cada consulta. A estrutura oferece suporte à recuperação composta em várias bases de conhecimento com reclassificação para maior precisão. Para obter mais informações, consulte a documentação do LlamaIndex RAG.
- Observabilidade e avaliação — LlamaIndex integra-se a uma variedade de ferramentas de observabilidade e avaliação. Esse recurso de integração ajuda você a rastrear e depurar seus aplicativos, avaliar seu desempenho e monitorar os custos. Para obter mais informações, consulte a documentação de rastreamento, depuração e avaliação. LlamaIndex

Quando utilizar LlamaIndex

LlamaIndexé particularmente adequado para cenários de IA agentes que enfatizam fluxos de trabalho intensivos em dados e gerenciamento de conhecimento:

- Aplicativos com muitos documentos que exigem que os agentes processem, analisem e extraiam insights de grandes volumes de documentos corporativos, como contratos, relatórios, manuais e registros regulatórios
- Prototipagem rápida para cenários de produção em que as organizações desejam criar e implantar rapidamente agentes centrados em documentos sem sobrecarga de gerenciamento de infraestrutura
- Arquiteturas pioneiras que priorizam a precisão da recuperação e a relevância do contexto, especialmente ao trabalhar com documentos complexos e multimodais contendo tabelas, imagens e dados estruturados
- Fluxos de trabalho de documentos multiagentes que exigem agentes especializados para diferentes aspectos do processamento de documentos, como análise, análise, resumo e verificação de conformidade

Abordagem de implementação para LlamaIndex

LlamaIndex fornece blocos de construção de baixo nível e abstrações de alto nível que acomodam diferentes abordagens de implementação:

- Desenvolvimento rápido de aplicativos RAG funcionais em apenas algumas linhas de código usando LlamaIndex alto nível APIs. Essa abordagem torna LlamaIndex acessível para equipes de negócios e desenvolvedores que são novos na IA agente.
- Integração empresarial LlamaHub por meio de sistemas corporativos populares SharePoint, incluindo Amazon Simple Storage Service (Amazon S3), bancos de dados e. APIs Essa abordagem permite uma integração perfeita com a infraestrutura de dados existente.
- Opções flexíveis de implantação entre implantações auto-hospedadas de código aberto para controle máximo ou serviços LlamaCloud gerenciados para reduzir a sobrecarga operacional e os recursos corporativos.
- Os aplicativos podem começar com mecanismos de consulta simples e adicionar progressivamente recursos agentes, orquestração de vários agentes e fluxos de trabalho complexos à medida que os requisitos evoluem.

Exemplo real de LlamaIndex

Este exemplo se concentra em uma subsidiária de uma empresa aeroespacial especializada em soluções de navegação e operações de aviação. Eles precisam enfrentar um desafio crescente que

envolve pilotar testes descoordenados de chatbots de IA. Os testes resultaram em trabalho repetido, longos ciclos de desenvolvimento, obstáculos de conformidade e implementações isoladas em toda a organização.

Eles desenvolveram uma estrutura de agente unificada, uma solução reutilizável baseada em modelos criada na estrutura de LlamaIndex código aberto que torna a criação de agentes muito mais eficiente. Eles compararam várias estruturas concorrentes, tanto orientadas por cadeias quanto baseadas em gráficos. Por fim, eles LlamaIndex escolheram três vantagens essenciais: seu design flexível, componentes modulares e controles de orquestração prontos para produção.

A plataforma reduz o tempo de desenvolvimento e implantação do agente em 87%, de 512 para 64 horas. Essa redução foi alcançada ao permitir que as equipes criassem agentes com aproximadamente 50 linhas de código e um arquivo de configuração JSON. As equipes utilizaram uma estrutura unificada com segurança integrada, conformidade e acesso privilegiado ao sistema. Para obter mais detalhes, consulte os estudos de caso de LlamaIndex clientes.

Comparando estruturas de IA agênticas

Ao selecionar uma estrutura de IA agente para o desenvolvimento de agentes autônomos, considere como cada opção se alinha aos seus requisitos específicos. Considere não apenas suas capacidades técnicas, mas também sua adequação organizacional, incluindo a experiência da equipe, a infraestrutura existente e os requisitos de manutenção de longo prazo. Muitas organizações podem se beneficiar de uma abordagem híbrida, aproveitando várias estruturas para diferentes componentes de seu ecossistema autônomo de IA.

A tabela a seguir compara os níveis de maturidade (mais forte, forte, adequado ou fraco) de cada estrutura nas principais dimensões técnicas. Para cada estrutura, a tabela também inclui informações sobre as opções de implantação de produção e a complexidade da curva de aprendizado.

Framework	AWS integração	Suporte autônomo a vários agentes	complexidade do fluxo de trabalho autônomo	Capacidades multimodais	Seleção do modelo de fundação	Integração da API LLM	Implantação de produção	Curva de aprendizado
AutoGen	Fraco	Forte	Forte	Adequado	Adequado	Forte	Faça você	Íngreme

							mesmo (DIY)	
CrewAI	Fraco	Forte	Adequado	Fraco	Adequado	Adequado	FAÇA VOCÊ MESMO	Moderada
LangChain / LangGraph	Adequado	Forte	Mais forte	Mais forte	Mais forte	Mais forte	Plataforma ou faça você mesmo	Íngreme
LlamaIndex	Adequado	Adequado	Forte	Adequado	Forte	Forte	Plataforma ou faça você mesmo	Moderada
Strands Agents	Mais forte	Forte	Mais forte	Forte	Forte	Mais forte	FAÇA VOCÊ MESMO	Moderada

Considerações na escolha de uma estrutura de IA agente

Ao desenvolver agentes autônomos, considere os seguintes fatores-chave:

- AWS integração de infraestrutura — As organizações em que investem fortemente se AWS beneficiarão mais das integrações nativas do Strands Agents with Serviços da AWS para fluxos de trabalho autônomos. Para obter mais informações, consulte Resumo AWS semanal (AWS blog).
- Seleção do modelo de fundação — Considere qual estrutura fornece o melhor suporte para seus modelos de fundação preferidos (por exemplo, modelos Amazon Nova no Amazon Bedrock ou Anthropic Claude), com base nos requisitos de raciocínio do seu agente autônomo. Para obter mais informações, consulte Criação de agentes eficazes no Anthropic site.
- Integração da API LLM — Avalie as estruturas com base em sua integração com suas interfaces de serviço preferidas de modelo de linguagem grande (LLM) (por exemplo, Amazon Bedrock

ouOpenAI) para implantação em produção. Para obter mais informações, consulte Interfaces de modelo na Strands Agents documentação.

- Requisitos multimodais — Para agentes autônomos que precisam processar texto, imagens e fala, considere os recursos multimodais de cada estrutura. Para obter mais informações, consulte Multimodalidade na LangChain documentação.
- Complexidade do fluxo de trabalho autônomo — Fluxos de trabalho autônomos mais complexos com gerenciamento de estado sofisticado podem favorecer os recursos avançados da máquina de estado. de. LangGraph
- Colaboração autônoma em equipe — Projetos que exigem colaboração autônoma explícita baseada em funções entre agentes especializados podem se beneficiar da arquitetura orientada à equipe do. CrewAI
- Paradigma de desenvolvimento autônomo — equipes que preferem padrões conversacionais e assíncronos para agentes autônomos podem preferir a arquitetura orientada a eventos do. AutoGen
- Abordagem gerenciada ou baseada em código — Organizações que desejam uma experiência totalmente gerenciada com o mínimo de codificação devem considerar os Amazon Bedrock Agents. Organizações que exigem uma personalização mais profunda podem preferir Strands Agents outras estruturas com recursos especializados que se alinhem melhor aos requisitos específicos de agentes autônomos.
- Prontidão de produção para sistemas autônomos — considere as opções de implantação, os recursos de monitoramento e os recursos corporativos para agentes autônomos de produção.

Plataformas

As plataformas de IA da Agentic fornecem as camadas básicas de tempo de execução, orquestração e integração necessárias para implantar, escalar e gerenciar sistemas agentes de nível de produção. As estruturas definem como os agentes são criados e os protocolos governam a forma como eles se comunicam. As plataformas fornecem o ambiente em que esses agentes operam, colaboram e evoluem com segurança em grande escala.

As plataformas Agentic combinam recursos de execução de modelos, gerenciamento de contexto, integração de ferramentas, observabilidade e governança em ambientes unificados. Essas plataformas permitem que as organizações passem da experimentação para a implantação em escala corporativa.

Nesta seção:

Por que as plataformas são importantes

As plataformas de IA da Agentic são essenciais para organizações que buscam operacionalizar sistemas autônomos na produção. Eles oferecem os seguintes recursos:

- Forneça orquestração de tempo de execução para hospedagem, escalabilidade e coordenação de agentes.
- Gerencie o estado, o contexto e a memória em fluxos de trabalho de vários agentes.
- Ofereça controles de segurança, identidade e governança alinhados aos padrões corporativos.
- Integre-se com ecossistemas de ferramentas e sistemas externos por meio de padrões APIs ou protocolos.
- Permita a observabilidade e a auditabilidade nas interações dos agentes e nos fluxos de eventos.
- Support a interoperabilidade entre modelos, permitindo que os agentes usem vários modelos básicos em um único ambiente.

Esses recursos transformam agentes individuais em sistemas coordenados e adaptáveis que podem operar de forma confiável dentro dos limites corporativos e regulatórios.

Tipos de plataformas de IA agênticas

As plataformas de IA da Agentic geralmente se enquadram em uma ou mais das seguintes categorias:

- Agente gerenciado — plataformas totalmente gerenciadas fornecem recursos integrados de infraestrutura, memória e orquestração. Eles reduzem a sobrecarga operacional e aceleram o tempo de produção.
- Orquestração de código aberto — As plataformas agentes de código aberto oferecem flexibilidade e transparência para organizações que preferem ambientes personalizáveis ou implantação local.
- Empresa híbrida — As plataformas híbridas integram componentes gerenciados e auto-hospedados, combinando a escalabilidade dos serviços gerenciados na nuvem com o controle dos sistemas corporativos.

Considerações sobre a seleção da plataforma

Ao selecionar ou projetar uma plataforma de IA agente, as organizações devem considerar o seguinte:

- Profundidade de integração — Avalie o quão bem a plataforma se integra às fontes de dados, ferramentas e protocolos existentes.
- Escalabilidade — Garanta que a plataforma possa ser escalada dinamicamente para suportar cargas de trabalho autônomas e colaboração entre vários agentes.
- Segurança e conformidade — avalie os recursos de privacidade, criptografia e governança de dados em relação aos requisitos organizacionais e regionais.
- Extensibilidade — escolha plataformas com arquiteturas modulares que permitam que novas ferramentas, modelos ou agentes sejam adicionados ao longo do tempo.
- Observabilidade — prefira plataformas que forneçam registros detalhados de telemetria, rastreabilidade e auditoria para interações com agentes.
- Eficiência de custos — considere modelos sem servidor ou baseados em uso para otimizar o custo de cargas de trabalho variáveis.

Amazon Bedrock Agents

O Amazon Bedrock Agents é um serviço totalmente gerenciado que permite criar e configurar agentes autônomos em seus aplicativos. Ele pode orquestrar interações entre modelos básicos, fontes de dados, aplicativos de software e conversas com usuários. Sua abordagem simplificada para criar agentes não exige que você provisione capacidade, gerencie a infraestrutura ou escreva código personalizado.

Principais características dos Amazon Bedrock Agents

O Amazon Bedrock Agents inclui os seguintes recursos principais:

- Serviço totalmente gerenciado — gerenciamento completo da infraestrutura sem a necessidade de provisionar capacidade ou gerenciar sistemas subjacentes. Para obter mais informações, consulte Automatizar tarefas em seu aplicativo usando agentes de IA na documentação do Amazon Bedrock.
- Desenvolvimento orientado por API — defina e execute agentes por meio de chamadas de API simples, especificando modelos, instruções, ferramentas e parâmetros de configuração. Para obter mais informações, consulte Criar e configurar o agente manualmente na documentação do Amazon Bedrock.
- Grupos de ação — defina ações específicas que seu agente pode realizar criando grupos de ação com esquemas de API. Para obter mais informações, consulte Usar grupos de ação para definir ações para seu agente realizar na documentação do Amazon Bedrock.
- Integração da base de conhecimento — Conecte-se perfeitamente às bases de conhecimento Amazon Bedrock para aumentar as respostas dos agentes com os dados da sua organização. Para obter mais informações, consulte Aumentar a geração de respostas para seu agente com base de conhecimento na documentação do Amazon Bedrock.
- Modelos de solicitação avançados — personalize o comportamento do agente por meio de modelos de solicitação para pré-processamento, orquestração, geração de respostas da base de conhecimento e pós-processamento. Para obter mais informações, consulte Melhorar a precisão do agente usando modelos avançados de solicitação no Amazon Bedrock na documentação do Amazon Bedrock.
- Rastreamento e observabilidade — acompanhe o processo de step-by-step raciocínio do agente usando recursos de rastreamento integrados. Para obter mais informações, consulte Rastrear o processo de step-by-step raciocínio do agente usando trace na documentação do Amazon Bedrock.

- Controle de versão e aliases — Crie várias versões do seu agente e implante-as por meio de aliases para lançamentos controlados. Para obter mais informações, consulte Implantar e usar um agente do Amazon Bedrock em seu aplicativo na documentação do Amazon Bedrock.

Quando usar o Amazon Bedrock Agents

O Amazon Bedrock Agents é particularmente adequado para cenários de agentes autônomos, incluindo:

- Organizações que desejam uma experiência totalmente gerenciada para criar e implantar agentes sem gerenciar a infraestrutura
- Projetos que exigem rápido desenvolvimento e implantação de agentes por meio de configuração em vez de código
- Casos de uso que se beneficiam da forte integração com outros recursos do Amazon Bedrock, como bases de conhecimento e guardrails
- Equipes sem recursos internos para criar agentes do zero, mas precisam de recursos autônomos prontos para produção

Abordagem de implementação para Amazon Bedrock Agents

O Amazon Bedrock Agents fornece uma abordagem de implementação baseada em configuração para as partes interessadas da empresa. O serviço permite que as organizações:

- Defina agentes por meio de chamadas de API Console de gerenciamento da AWS ou de chamadas sem escrever código complexo.
- Crie grupos de ação que especifiquem as operações APIs e que o agente pode realizar.
- Conecte as bases de conhecimento para fornecer informações específicas do domínio ao agente.
- Teste e repita o comportamento do agente por meio de uma interface visual.

Essa abordagem gerenciada permite que as equipes de negócios desenvolvam e implantem rapidamente agentes autônomos sem profundo conhecimento técnico no desenvolvimento de modelos de IA ou no gerenciamento de infraestrutura.

Exemplo real dos Amazon Bedrock Agents

Uma solução de operações financeiras (FinOps) descrita nesta postagem do AWS blog usa a estrutura multiagente Amazon Bedrock para criar um assistente de gerenciamento de custos na nuvem orientado por IA. O modelo econômico da Amazon Nova Foundation capacita a solução em que um agente FinOps supervisor central delega tarefas a agentes especializados. Esses agentes buscam e analisam dados de AWS gastos usando AWS Cost Explorer e geram recomendações de economia de custos usando. AWS Trusted Advisor

O sistema inclui acesso seguro do usuário por meio do Amazon Cognito, um front-end hospedado no AWS Amplify, e grupos de AWS Lambda ação para análise e previsão em tempo real. As equipes financeiras podem fazer perguntas em linguagem natural, como “Quais foram meus custos em fevereiro de 2025?” O sistema responde com detalhamentos, sugestões de otimização e previsões, tudo em uma arquitetura escalável e sem servidor implantada usando. AWS CloudFormation

Amazon Bedrock AgentCore

O Amazon Bedrock AgentCore é uma plataforma agente para criar, implantar e operar agentes altamente capazes com segurança em grande escala usando qualquer estrutura, modelo ou protocolo. Usando AgentCore, você pode fazer o seguinte, tudo sem nenhum gerenciamento de infraestrutura:

- Crie agentes com mais rapidez.
- Permita que os agentes realizem ações em todas as ferramentas e dados.
- Execute agentes de forma segura com baixa latência e tempos de execução estendidos.
- Monitore os agentes na produção.

AgentCore elimina o trabalho pesado indiferenciado de criar uma infraestrutura especializada para agentes, permitindo que você acelere a produção de seus agentes. Seus serviços podem ser usados juntos ou de forma independente e são compatíveis com qualquer estrutura CrewAILangGraph, incluindoLlamaIndex,, Strands Agents e. AgentCoretambém é compatível com qualquer modelo de base disponível dentro ou fora do Amazon Bedrock, oferecendo flexibilidade máxima.

AgentCore é composto por vários serviços principais:

- Amazon Bedrock AgentCore Runtime — Fornece um ambiente seguro, sem servidor e escalável para hospedar e executar seus agentes, sem precisar gerenciar qualquer infraestrutura necessária para implantar e executar agentes ou ferramentas de IA.
- Amazon Bedrock AgentCore Memory — oferece um sistema de memória gerenciada, permitindo que os agentes retenham o contexto das interações para conversas mais personalizadas e coerentes, mantendo o conhecimento imediato e de longo prazo.
- Amazon Bedrock AgentCore Gateway — Simplifica o processo de criação, proteção e localização das ferramentas certas para agentes. Com o AgentCore Gateway, os desenvolvedores podem converter APIs funções Lambda e serviços existentes em ferramentas compatíveis com o Model Context Protocol (MCP) e disponibilizá-las aos agentes.
- Amazon Bedrock AgentCore Identity — fornece um serviço de gerenciamento de identidade e acesso de agentes seguro e escalável que acelera o desenvolvimento de agentes de IA. Com o AgentCore Identity, você pode atribuir identidades exclusivas e verificáveis aos agentes, permitindo um controle de acesso refinado e interações seguras baseadas em agentes com sistemas corporativos.
- Ferramentas AgentCore integradas do Amazon Bedrock — Permite que você use ferramentas integradas para aprimorar seu fluxo de trabalho de desenvolvimento e teste. Use essas ferramentas para interagir com seu aplicativo de forma eficaz, permitindo que agentes de IA escrevam e executem códigos com segurança em ambientes de sandbox. Use a ferramenta do navegador para permitir que agentes de IA interajam com sites em grande escala.
- Amazon Bedrock AgentCore Observability — fornece recursos de registro e monitoramento, oferecendo visibilidade em tempo real do desempenho e do comportamento do seu agente para facilitar a depuração e a otimização.

Principais características do AgentCore

AgentCore inclui os seguintes recursos principais:

- Totalmente gerenciado e extensível — AgentCore é um serviço totalmente gerenciado, o que significa que AWS lida com a infraestrutura e a manutenção subjacentes. Também é extensível, o que permite personalizar e aprimorar a funcionalidade de seus agentes. Para obter mais informações, consulte Introdução ao AgentCore Runtime na AgentCore documentação.
- Memória de longo e curto prazo — Ofereça interações mais personalizadas e relevantes equipando os agentes com um sistema de memória para relembrar o contexto das conversas

atuais e do conhecimento de longo prazo. Para obter mais informações, consulte Introdução à AgentCore memória na AgentCore documentação.

- Desenvolvimento e integração simplificados de ferramentas — Permita que seus agentes descubram e usem ferramentas por meio de um único endpoint seguro. Transforme rapidamente seus recursos corporativos existentes em ferramentas prontas para agentes com apenas algumas linhas de código, liberando os desenvolvedores para se concentrarem na criação de recursos exclusivos. Para obter mais informações, consulte Introdução ao AgentCore Gateway na AgentCore documentação.
- Infraestrutura segura e escalável — AgentCore fornece um ambiente seguro e escalável para a implantação e operação de agentes. Ele inclui recursos para gerenciamento de identidade e acesso, criptografia de dados e segurança de rede. Para obter mais informações, consulte Introdução ao AgentCore Identity na AgentCore documentação.
- Integração com uma ampla variedade de ferramentas — Permite que você integre seus agentes a uma variedade de ferramentas, incluindo um interpretador de código e uma ferramenta de navegador que você pode criar usando as ferramentas AgentCore integradas. Para obter mais informações, consulte Começar com o AgentCore Code Interpreter e Começar com o AgentCore Browser na AgentCore documentação.
- Observabilidade e monitoramento abrangentes — Obtenha visibilidade profunda de seus agentes com ferramentas abrangentes para rastrear, depurar e monitorar seu desempenho na produção. Visualize todo o caminho de execução do agente para auditar seu raciocínio e resolver falhas. Use painéis em tempo real e dados de telemetria padronizados para rastrear as principais métricas operacionais. Para obter mais informações, consulte Adicionar observabilidade aos seus AgentCore recursos do Amazon Bedrock na AgentCore documentação.

Quando usar AgentCore

AgentCore é particularmente adequado para cenários de agentes autônomos, incluindo:

- Organizações que desejam acelerar o desenvolvimento e reduzir a sobrecarga operacional com um serviço totalmente gerenciado que lida com infraestrutura, segurança, ferramentas integradas, observabilidade e escalabilidade
- Projetos que precisam de flexibilidade com serviços modulares que funcionam juntos ou de forma independente e são compatíveis com qualquer estrutura, como CrewAI ouLangGraph, e qualquer modelo básico de qualquer fonte

- Casos de uso que exigem agentes interativos e conversacionais que precisam manter o contexto e aprender com as interações anteriores para fornecer respostas personalizadas e relevantes
- Agentes habilitados para realizar tarefas complexas por meio de integração simples com diversos aplicativos, fontes de dados e APIs

Abordagem de implementação para AgentCore

AgentCore foi projetado para organizações que desejam transferir agentes de IA da prova de conceito, criada usando estruturas de agentes de código aberto ou personalizadas, para a produção. Com AgentCore, as organizações podem fazer o seguinte:

- Implante agentes com segurança em uma infraestrutura sem servidor, oferecendo suporte a qualquer estrutura e modelo, com isolamento de sessão e gerenciamento integrado de identidade e acesso para end-to-end segurança e conformidade. Crie rapidamente agentes AgentCore Runtime para as principais estruturas de agentes usando o kit de ferramentas inicial.
- Melhore os agentes integrando memória persistente para retenção de contexto, simplificando o desenvolvimento e a integração de ferramentas por meio AgentCore do Gateway. Aproveite a ferramenta de navegador e o interpretador de código integrados para fluxos de trabalho avançados.
- Rastreie, depure e monitore agentes de IA em produção usando painéis de observabilidade desenvolvidos pelo Amazon CloudWatch Application Insights e OpenTelemetry rastreando as principais métricas dos AgentCore recursos (tempo de execução, memória, gateway e ferramentas).
- Acelere a implantação e a inovação com serviços modulares totalmente gerenciados, blocos compostos juntos ou de forma independente, com qualquer fornecedor de modelos e estruturas de agentes. Essa flexibilidade ajuda as organizações a passar do protótipo para a produção com mais rapidez.

Essa abordagem gerenciada permite que as organizações criem, implantem e executem com rapidez e segurança agentes de IA e sistemas multiagentes de nível empresarial em qualquer escala.

Exemplo real de AgentCore

AWS observou que um dos maiores bancos da América Latina usa AI/ML há anos uma experiência bancária digital hiperpersonalizada e segura. O banco está expandindo os serviços de inteligência artificial AgentCore para fornecer aos clientes interações intuitivas, segurança aprimorada e maior

automação. De acordo com o CTO, AgentCore espera-se que apoie seus esforços para cumprir os compromissos dos clientes em grande escala. AgentCore fornece aos desenvolvedores as ferramentas e a flexibilidade para criar e gerenciar agentes, ao mesmo tempo em que ajuda a garantir a conformidade com as regulamentações financeiras.

Protocolos

Os agentes de IA exigem protocolos de comunicação padronizados para interagir com outros agentes e serviços. As organizações que implementam arquiteturas de agentes enfrentam desafios significativos em relação à interoperabilidade, independência de fornecedores e à preparação de seus investimentos para o futuro.

Esta seção ajuda você a navegar pelo cenário de agent-to-agent protocolos com foco em padrões abertos que maximizam a flexibilidade e a interoperabilidade. (Para obter informações sobre agent-to-tool protocolos, consulte Estratégia de integração de ferramentas mais adiante neste guia.)

Esta seção destaca o Model Context Protocol (MCP), um padrão aberto originalmente desenvolvido Anthropic em 2024. Hoje, apoia AWS ativamente o MCP por meio de contribuições para o desenvolvimento e implementação do protocolo. AWS está colaborando com as principais estruturas de agentes de código aberto, incluindoLangGraph, e CrewAILlamaIndex, para moldar o futuro da comunicação entre agentes no protocolo. Para obter mais informações, consulte Protocolos abertos para interoperabilidade de agentes, parte 1: Comunicação entre agentes no MCP (blog).AWS

Nesta seção:

- Por que a seleção de protocolos é importante
- Agent-to-agent protocolos
- Seleção de protocolos agentes
- Estratégia de implementação para protocolos agentes
- Começando com o MCP
- ???

Por que a seleção de protocolos é importante

A seleção de protocolos molda fundamentalmente como você pode criar e desenvolver sua arquitetura de agente de IA. Ao escolher protocolos que oferecem suporte à portabilidade entre estruturas de agentes, você ganha a flexibilidade de combinar diferentes sistemas e fluxos de trabalho de agentes para atender às suas necessidades específicas.

Os protocolos abertos permitem que você integre agentes em várias estruturas. Por exemplo, use LangChain para prototipagem rápida e implemente sistemas de produção com comunicação por

meio de um protocolo comumStrands Agents, como MCP ou o protocolo Agent2Agent (A2A). Essa flexibilidade reduz a dependência de provedores específicos de IA, simplifica a integração com os sistemas existentes e permite que você aprimore os recursos dos agentes ao longo do tempo.

Protocolos bem projetados também estabelecem padrões de segurança consistentes para autenticação e autorização em todo o ecossistema de agentes. Mais importante ainda, a portabilidade do protocolo preserva sua liberdade de adotar novas estruturas e recursos de agentes à medida que eles surgem. A escolha de protocolos abertos protege seu investimento no desenvolvimento de agentes e, ao mesmo tempo, mantém a interoperabilidade com sistemas de terceiros.

Vantagens dos protocolos abertos

Ao implementar suas próprias extensões ou criar sistemas de agentes personalizados, os protocolos abertos oferecem vantagens convincentes:

- Documentação e transparência — Normalmente fornecem documentação abrangente e implementações transparentes
- Suporte comunitário — Acesso a comunidades mais amplas de desenvolvedores para solução de problemas e melhores práticas
- Garantias de interoperabilidade — Melhor garantia de que suas extensões funcionarão em diferentes implementações
- Compatibilidade futura — risco reduzido de falhas, alterações ou descontinuação
- Influência no desenvolvimento — Oportunidade de contribuir para a evolução do protocolo

Agent-to-agent protocolos

A tabela a seguir fornece uma visão geral dos protocolos de agentes que permitem que vários agentes colaborem, deleguem tarefas e compartilhem informações.

Protocolo	Ideal para	Considerações
Comunicação interagente MCP	Organizações que buscam padrões flexíveis de colaboração de agentes	• Uma extensão do Model Context Protocol (MCP) proposta por AWS that se baseia em sua base

		existente para comunicação agent-to-agent • Permite a colaboração perfeita dos agentes com segurança OAuth baseada
Protocolo A2A	Ecossistemas de agentes multiplataforma	• Apoiado por Google • Padrão mais novo com adoção mais limitada em comparação com o MCP

Decidindo entre as opções de protocolo

Ao implementar a agent-to-agent comunicação, combine seus requisitos específicos de comunicação com os recursos de protocolo apropriados. Padrões de interação diferentes exigem recursos de protocolo diferentes. A tabela a seguir descreve os padrões comuns de comunicação e recomenda as opções de protocolo mais adequadas para cada cenário.

Padrão	Descrição	Escolha de protocolo ideal
Solicitação e resposta simples	Interações pontuais entre agentes	MCP com fluxos sem estado
Diálogos estatais	Conversas contínuas com contexto	MCP com gerenciamento de sessões
Colaboração com vários agentes	Interações complexas entre vários agentes	Interagente MCP ou AutoGen
Fluxos de trabalho baseados em equipe	Equipes hierárquicas de agentes com funções definidas	Interagente MCP, ou CrewAI AutoGen

Além dos padrões de comunicação, vários fatores técnicos e organizacionais podem influenciar sua seleção de protocolos. A tabela a seguir descreve as principais considerações que podem ajudá-lo a avaliar qual protocolo está mais alinhado com seus requisitos específicos de implementação.

Consideração	Descrição	Exemplo
Modelo de segurança	Requisitos de autenticação e autorização	OAuth 2.0 em MCP
Ambiente de implantação	Onde os agentes correrão e se comunicarão	Máquina distribuída ou única
Compatibilidade com ecossistemas	Integração com estruturas de agentes existentes	LangChain ou Strands Agents
Necessidades de escalabilidade	Crescimento esperado nas interações entre agentes	Capacidades de streaming do MCP

Seleção de protocolos agentes

Para a maioria das organizações que criam sistemas de agentes de produção, o Model Context Protocol (MCP) oferece a base mais abrangente e bem apoiada para agent-to-agent comunicação. O MCP se beneficia das contribuições ativas de desenvolvimento AWS e da comunidade de código aberto.

Selecionar os protocolos agentes corretos é importante para organizações que desejam implementar a IA agente de forma eficaz. As considerações diferem com base no contexto organizacional.

Considerações sobre a seleção de protocolos agentes

As organizações devem considerar as seguintes melhores práticas ao selecionar protocolos para sistemas de IA agentes:

- Priorize padrões abertos — As organizações devem adotar protocolos abertos, como o MCP, para ajudar a garantir a interoperabilidade e a extensibilidade a longo prazo e reduzir o risco de dependência de fornecedores.
- Equilibre velocidade e flexibilidade — As startups e os primeiros usuários podem começar com protocolos proprietários bem suportados para um rápido desenvolvimento, mas devem definir um caminho de migração para padrões abertos à medida que os sistemas amadurecem.

- Implemente camadas de abstração — As empresas devem implementar a abstração de protocolos para simplificar a migração, permitir a adoção híbrida e estratégias de integração preparadas para o futuro.
- Enfatize a segurança e a conformidade — Organizações em setores regulamentados devem selecionar protocolos com recursos robustos de autenticação, criptografia e auditoria para atender aos requisitos de governança e conformidade.
- Avalie a maturidade do ecossistema — Todas as organizações devem avaliar a saúde, a adoção e o apoio comunitário de cada protocolo para garantir a sustentabilidade e minimizar a dívida técnica.
- Envolva-se no desenvolvimento de padrões — As organizações devem participar de órgãos de padrões ou comunidades de código aberto para ajudar a moldar a evolução do protocolo e influenciar as melhores práticas.
- Leve em conta a soberania dos dados — O governo e os setores regulamentados devem garantir que as opções de protocolo estejam alinhadas aos requisitos de residência e soberania dos dados em todas as regiões de implantação.
- Aproveite os serviços gerenciados — sempre que possível, use implementações gerenciadas ou sem servidor de protocolos agentes para reduzir a complexidade operacional e acelerar a implantação.

Estratégia de implementação para protocolos agentes

Para implementar protocolos agentes de forma eficaz em sua organização, considere as seguintes etapas estratégicas:

1. Comece com o alinhamento de padrões — adote protocolos abertos estabelecidos sempre que possível.
2. Crie camadas de abstração — Implemente adaptadores entre seus sistemas e protocolos específicos.
3. Contribua com padrões abertos — participe de comunidades de desenvolvimento de protocolos.
4. Monitore a evolução do protocolo — mantenha-se informado sobre os padrões e atualizações emergentes.
5. Teste a interoperabilidade regularmente — verifique se suas implementações permanecem compatíveis.

Começando com o MCP

AWS apoia ativamente o Model Context Protocol (MCP) por meio de contribuições para o desenvolvimento e implementação do protocolo. AWS está colaborando com as principais estruturas de agentes de código aberto, incluindoLangGraph, e CrewAILlamaIndex, para moldar o futuro da comunicação entre agentes no protocolo.

Para implementar o MCP em sua arquitetura de agente, execute as seguintes ações:

1. Explore as implementações do MCP em estruturas como o SDK. Strands Agents
2. Revise a documentação técnica do Model Context Protocol.
3. Leia Protocolos abertos para interoperabilidade de agentes, parte 1: comunicação entre agentes no MCP (AWS blog) para saber mais sobre a interoperabilidade de agentes.
4. Junte-se à comunidade MCP para influenciar a evolução do protocolo.

O MCP fornece uma camada de comunicação que permite que os agentes interajam com dados e serviços externos e também pode ser usada para permitir que os agentes interajam com outros agentes. A implementação de transporte HTTP Streamable do protocolo oferece aos desenvolvedores um conjunto abrangente de padrões de interação sem precisar reinventar a roda. Esses padrões oferecem suporte a request/response fluxos sem estado e ao gerenciamento de sessões com monitoramento de estado com persistência. IDs

Ao adotar protocolos abertos como o MCP, você posiciona sua organização para criar sistemas de agentes que permaneçam flexíveis, interoperáveis e adaptáveis à medida que a tecnologia de IA evolui. Para obter informações sobre a implementação do agent-to-tool protocolo, consulte Estratégia de integração de ferramentas mais adiante neste guia.

Começando com o A2A

O protocolo Agent2Agent (A2A) permite a colaboração descentralizada entre agentes por meio de uma camada semântica compartilhada. Em vez de rotear todo o trabalho por meio de um orquestrador central, o A2A permite que os agentes se descubram, anunciem suas capacidades, negociem tarefas e compartilhem contexto usando um protocolo leve baseado em JSON. Cada agente publica um manifesto de capacidade.

O exemplo a seguir mostra um manifesto simplificado da capacidade A2A que anuncia as ações suportadas, as entradas necessárias e os metadados operacionais de um agente para permitir a descoberta e a negociação de tarefas:

```
{
  "can": ["summarize.text", "extract.keywords"],
  "needs": ["document.input"],
  "meta": { "version": "1.0.3", "latencyMs": 120 }
}
```

Esse modelo permite a correspondência dinâmica de capacidades, a delegação no meio da tarefa e a colaboração entre organizações. Os agentes podem se auto-organizar em torno de tarefas, formar grupos de trabalho temporários e se adaptar à medida que novos recursos entram ou saem do sistema.

O A2A suporta interações que vão desde simples solicitações sem estado até sessões de negociação em várias etapas, incluindo:

- peer-to-peerMensagens diretas para colaboração de baixa latência
- Negociação semântica de tarefas, em que os agentes selecionam o par mais adequado
- Descoberta baseada em capacidades, permitindo a divisão emergente do trabalho
- Ancoragem de sessão para interações com estado em várias etapas

Ao adotar protocolos abertos e nativos de agentes, como o A2A, as organizações criam sistemas de IA que são modulares, interoperáveis e capazes de colaboração transfronteiriça. O A2A garante que os ecossistemas de agentes permaneçam flexíveis e possam evoluir à medida que novos agentes, equipes ou sistemas externos são introduzidos, sem exigir camadas rígidas de orquestração ou acoplamento prévio.

Para implementar o protocolo A2A em sua arquitetura de agente, execute as seguintes ações:

1. Analise a especificação do protocolo A2A — Leia a versão mais recente da especificação do protocolo Agent2Agent (A2A) para saber como os manifestos de capacidade, os fluxos de negociação e o handshake do agente operam.
2. Explore tempos de execução compatíveis com A2A — Avalie estruturas como o SDK Strands Agents ou camadas de tempo de execução personalizadas que oferecem suporte a manifestos e negociações de recursos no estilo A2A. peer-to-peer

3. Implemente um manifesto de capacidades para seus agentes — defina os `meta` campos e os campos de `can` cada agente para permitir a descoberta, a combinação e a colaboração em nível de intenção. `needs`
4. Experimente os padrões de negociação A2A — use o ciclo de solicitação-oferta—aceitação, consultas de recursos estruturados ou descobertas baseadas em fofocas para entender como os agentes raciocinam sobre quem deve lidar com uma tarefa.
5. Teste o A2A em um ambiente de infraestrutura mista — Combine a negociação entre pares A2A com o roteamento de eventos nativo da AWS Amazon para avaliar padrões de coordenação híbrida. EventBridge
6. Junte-se à comunidade A2A — envolva-se com o grupo de trabalho aberto para se manter atualizado com extensões, recomendações de segurança e melhorias na interoperabilidade entre fornecedores e contribua para o desenvolvimento do protocolo.

Ferramentas

Os agentes de IA agregam valor ao interagir com ferramentas externas e fontes de dados para realizar tarefas úteis. APIs A estratégia certa de integração de ferramentas afeta diretamente as capacidades, a postura de segurança e a flexibilidade de longo prazo do seu agente.

Esta seção ajuda você a navegar pelo cenário de integração de ferramentas com foco em padrões abertos que maximizam sua liberdade e flexibilidade. A seção destaca o Model Context Protocol (MCP) para integração de ferramentas e analisa ferramentas específicas da estrutura e meta-ferramentas especializadas que aprimoram os fluxos de trabalho dos agentes.

Nesta seção:

- Categorias de ferramentas
- Ferramentas baseadas em protocolos
- Ferramentas nativas da estrutura
- Meta-ferramentas
- Estratégia de integração de ferramentas
- Melhores práticas de segurança para integração de ferramentas

Categorias de ferramentas

Os sistemas de agentes de construção envolvem três categorias principais de ferramentas.

Ferramentas baseadas em protocolos

As ferramentas baseadas em protocolos usam protocolos padronizados para agent-to-tool comunicação:

- Ferramentas MCP — ferramentas padrão abertas que funcionam em várias estruturas com opções de execução local e remota.
- OpenAIchamada de função — ferramentas proprietárias que são específicas para OpenAI modelos.
- Anthropictools — Uma variação da chamada de OpenAI função para ferramentas proprietárias que são específicas dos modelos Anthropic Claude.

Ferramentas nativas da estrutura

As ferramentas nativas da estrutura são incorporadas diretamente em estruturas específicas de agentes:

- Strands Agents ferramentas — quick-to-implement Ferramentas leves, específicas para a Strands Agents estrutura.
- LangChainferramentas — ferramentas Python baseadas em ferramentas que estão totalmente integradas ao LangChain ecossistema.
- LlamaIndexferramentas — Ferramentas que são otimizadas para recuperação e processamento de dados internosLlamaIndex.

Meta-ferramentas

As meta-ferramentas aprimoram os fluxos de trabalho dos agentes sem realizar ações externas diretamente:

- Ferramentas de fluxo de trabalho — gerencie o fluxo de execução do agente, a lógica de ramificação e o gerenciamento do estado.
- Ferramentas de gráfico de agentes — coordene vários agentes em fluxos de trabalho complexos.
- Ferramentas de memória — fornecem armazenamento e recuperação persistentes de informações em todas as sessões do agente.
- Ferramentas de reflexão — Permita que os agentes analisem e melhorem seu próprio desempenho.

Ferramentas baseadas em protocolos

Ao considerar ferramentas baseadas em protocolos, o Model Context Protocol (MCP) fornece a base mais abrangente e flexível para a integração de ferramentas. Conforme declarado na postagem do blog AWS Open Source sobre interoperabilidade de agentes, AWS adotou o MCP como um protocolo estratégico, contribuindo ativamente para seu desenvolvimento.

A tabela a seguir descreve as opções para a implantação da ferramenta MCP.

Modelo de implantação	Descrição	Ideal para	Implementação
Baseado em estúdio local	As ferramentas são executadas no mesmo processo que o agente	Desenvolvimento, teste e ferramentas simples	Rápido de implementar sem sobrecarga de rede
Baseado em eventos enviados pelo servidor local (SSE)	As ferramentas são executadas localmente, mas se comunicam por HTTP	Ferramentas locais mais complexas com separação de interesses	Melhor isolamento, mas ainda baixa latência
HTTP remoto que pode ser transmitido	Ferramentas executadas em servidores remotos	Ambientes de produção e ferramentas compartilhadas	Escalável e gerenciado centralmente

Os MCP oficiais SDKs estão disponíveis para criar ferramentas MCP:

- PythonSDK — Implementação abrangente com suporte total ao protocolo
- TypeScriptSDK —JavaScript/TypeScriptimplementação para aplicativos web
- JavaSDK — implementação Java para aplicativos corporativos

Eles SDKs fornecem os alicerces para a criação de ferramentas compatíveis com MCP em sua linguagem preferida, com implementações consistentes da especificação do protocolo.

Além disso, AWS implementou o MCP no Strands AgentsSDK. O Strands Agents SDK fornece uma maneira simples de criar e usar ferramentas compatíveis com MCP. Uma documentação abrangente está disponível no Strands Agents GitHub repositório. Para casos de uso mais simples ou ao trabalhar fora da Strands Agents estrutura, o MCP oficial SDKs oferece implementações diretas do protocolo em vários idiomas.

Recursos de segurança das ferramentas MCP

Os recursos de segurança das ferramentas MCP incluem o seguinte:

- OAuth Autenticação 2.0/2.1 — Autenticação padrão do setor

- Escopo de permissões — controle de acesso refinado para ferramentas
- Descoberta da capacidade da ferramenta — descoberta dinâmica das ferramentas disponíveis
- Tratamento estruturado de erros — padrões de erro consistentes

Introdução às ferramentas MCP

Para implementar o MCP para integração de ferramentas, execute as seguintes ações:

1. Explore o Strands AgentsSDK para uma implementação de MCP pronta para produção.
2. Revise a documentação técnica do MCP para entender os principais conceitos.
3. Use os exemplos práticos descritos nesta postagem do blog de código AWS aberto.
4. Comece com ferramentas locais simples antes de passar para ferramentas remotas.
5. Junte-se à comunidade MCP para influenciar a evolução do protocolo.

Conheça o AgentCore Gateway

O Amazon Bedrock AgentCore Gateway fornece uma maneira fácil e segura para os desenvolvedores criarem, implantarem, descobrirem e se conectarem às ferramentas de MCP e a outros endpoints de destino em grande escala. Com o AgentCore Gateway, os desenvolvedores podem converter APIs AWS Lambda funções e serviços existentes em ferramentas compatíveis com MCP. Então, com apenas algumas linhas de código, eles podem disponibilizar essas ferramentas aos agentes por meio de endpoints do AgentCore Gateway. AgentCore O Gateway suporta OpenAPISmithy, e Lambda como tipos de entrada e é a única solução que fornece autenticação abrangente de entrada e autenticação de saída em um serviço totalmente gerenciado.

Ferramentas nativas da estrutura

Embora o Model Context Protocol (MCP) forneça a base mais flexível, as ferramentas nativas da estrutura oferecem vantagens para casos de uso específicos.

O Strands AgentsSDK oferece ferramentas Python baseadas caracterizadas por seu design leve que requer sobrecarga mínima para operações simples. Eles permitem uma implementação rápida e permitem que os desenvolvedores criem ferramentas com apenas algumas linhas de código. Além disso, eles são totalmente integrados para funcionar perfeitamente dentro da Strands Agents estrutura.

O exemplo a seguir demonstra como criar uma ferramenta climática simples usandoStrands Agents. Os desenvolvedores podem transformar rapidamente Python funções em ferramentas acessíveis por agentes com o mínimo de sobrecarga de código e gerar automaticamente a documentação apropriada a partir da docstring da função.

```
#Example of a simple Strands native tool

@tool

def weather(location: str) -> str:

"""Get the current weather for a location""" #

Implementation here

return f"The weather in {location} is sunny."
```

Para prototipagem rápida ou casos de uso simples, as ferramentas nativas da estrutura podem acelerar o desenvolvimento. No entanto, para sistemas de produção, as ferramentas MCP oferecem melhor interoperabilidade e flexibilidade futura do que as ferramentas nativas da estrutura.

A tabela a seguir fornece uma visão geral de outras ferramentas específicas da estrutura.

Framework	Tipo de ferramenta	Vantagens	Considerações
AutoGen	Definições de funções	Forte suporte multiagente	Microsoftecossistema
LangChain	Pythonaulas	Grande ecossistema de ferramentas pré-construídas	Bloqueio de estrutura
LlamaIndex	Funções do Python	Otimizado para operações de dados	Limitado a LlamaIndex

Meta-ferramentas

As meta-ferramentas não interagem diretamente com sistemas externos. Em vez disso, eles aprimoram as capacidades dos agentes implementando padrões agentes. Esta seção aborda o fluxo de trabalho, o gráfico do agente e as meta-ferramentas de memória.

Meta-ferramentas de fluxo de trabalho

As meta-ferramentas de fluxo de trabalho gerenciam o fluxo de execução do agente:

- Gerenciamento de estado — mantenha o contexto em várias interações de agentes
- Lógica de ramificação — Habilite caminhos de execução condicional
- Mecanismos de repetição — Lide com falhas com estratégias sofisticadas de repetição

Exemplos de estruturas com meta-ferramentas de fluxo de trabalho incluem LangGraphrecursos de fluxo de trabalho. Strands Agents

Meta-ferramentas de gráficos de agentes

As meta-ferramentas de gráficos de agentes coordenam vários agentes trabalhando juntos:

- Delegação de tarefas — atribua subtarefas a agentes especializados
- Agregação de resultados — Combine resultados de vários agentes
- Resolução de conflitos — Resolva divergências entre agentes

Frameworks como AutoGene se CrewAIespecializam em coordenação gráfica de agentes.

Meta-ferramentas de memória

As meta-ferramentas de memória fornecem armazenamento e recuperação persistentes:

- Histórico de conversas — mantenha o contexto em todas as sessões
- Bases de conhecimento — Armazene e recupere informações específicas do domínio
- Armazenamentos vetoriais — Habilite recursos de pesquisa semântica

O sistema de recursos do MCP fornece uma maneira padronizada de implementar meta-ferramentas de memória que funcionam em diferentes estruturas de agentes.

Estratégia de integração de ferramentas

Sua escolha de estratégia de integração de ferramentas afeta diretamente o que seus agentes podem realizar e a facilidade com que seu sistema pode evoluir. Priorize protocolos abertos, como

o Model Context Protocol (MCP), enquanto usa estrategicamente ferramentas e meta-ferramentas nativas da estrutura. Dessa forma, você pode criar um ecossistema de ferramentas que permaneça flexível e poderoso à medida que a tecnologia de IA avança.

A seguinte abordagem estratégica para integração de ferramentas maximiza a flexibilidade e, ao mesmo tempo, atende às necessidades imediatas da sua organização:

1. Adote o MCP como sua base — O MCP fornece uma maneira padronizada de conectar agentes a ferramentas com recursos de segurança robustos. Comece com o MCP como seu protocolo de ferramenta principal para:
   - Ferramentas estratégicas que serão usadas em várias implementações de agentes.
   - Ferramentas sensíveis à segurança que exigem autenticação e autorização robustas.
   - Ferramentas que precisam de execução remota em ambientes de produção.
2. Use ferramentas nativas da estrutura quando apropriado — Considere as ferramentas nativas da estrutura para:
   - Prototipagem rápida durante o desenvolvimento inicial.
   - Ferramentas simples e não críticas com requisitos mínimos de segurança.
   - Funcionalidade específica da estrutura que aproveita recursos exclusivos.
3. Implemente meta-ferramentas para fluxos de trabalho complexos — Adicione meta-ferramentas para aprimorar sua arquitetura de agente:
   - Comece de forma simples com padrões básicos de fluxo de trabalho.
   - Adicione complexidade à medida que seus casos de uso amadurecem.
   - Padronize as interfaces entre agentes e meta-ferramentas.
4. Planeje a evolução — Crie com a flexibilidade futura em mente:
   - Documente as interfaces das ferramentas independentemente das implementações.
   - Crie camadas de abstração entre agentes e ferramentas.
   - Estabeleça caminhos de migração de protocolos proprietários para protocolos abertos.

Melhores práticas de segurança para integração de ferramentas

A integração de ferramentas afeta diretamente sua postura de segurança. Esta seção descreve as melhores práticas a serem consideradas em sua organização.

Autenticação e autorização

Use os seguintes controles de acesso robustos:

- Use OAuth 2.0/2.1 — Implemente a autenticação padrão do setor para ferramentas remotas.
- Implemente o menor privilégio — conceda às ferramentas somente as permissões de que elas precisam.
- Alterne as credenciais — atualize regularmente as chaves de API e os tokens de acesso.

Proteção de dados

Para ajudar a proteger os dados, adote as seguintes medidas:

- Valide entradas e saídas — Implemente a validação do esquema para todas as interações da ferramenta.
- Criptografe dados confidenciais — use o TLS para todas as comunicações remotas com ferramentas.
- Implemente a minimização de dados — passe somente as informações necessárias para as ferramentas.

Monitoramento e auditoria

Mantenha a visibilidade e o controle usando esses mecanismos:

- Registre todas as invocações de ferramentas — mantenha trilhas de auditoria abrangentes.
- Monitore anomalias — Detecte padrões incomuns de uso de ferramentas.
- Implemente a limitação de taxa — Evite abusos por meio de chamadas excessivas de ferramentas.

O modelo de segurança do Model Context Protocol (MCP) aborda essas preocupações de forma abrangente. Para obter mais informações, consulte Considerações de segurança na documentação do MCP.

Conclusão

O cenário da IA agêntica continua evoluindo rapidamente, oferecendo às organizações novas maneiras poderosas de criar sistemas inteligentes e autônomos. Este guia explorou três componentes essenciais para uma implementação bem-sucedida: estruturas que fornecem a base, plataformas que fornecem o ambiente, protocolos que permitem a comunicação e ferramentas que ampliam os recursos.

À medida que as estruturas amadurecem, você pode esperar maior interoperabilidade, padronização em torno de protocolos como o Model Context Protocol (MCP) e recursos de orquestração mais sofisticados para agentes autônomos. As organizações que estabelecem experiência com essas estruturas hoje estarão bem posicionadas para criar agentes cada vez mais autônomos e inteligentes que ofereçam um valor comercial significativo.

As plataformas fornecem o ambiente de execução, governança e ciclo de vida no qual os sistemas agentes operam. Eles lidam com questões como identidade, limites de segurança, observabilidade, gerenciamento de memória, aterramento de sessões e interação segura com ferramentas e dados. Em AWS ambientes, plataformas como tempos de execução de agentes gerenciados e serviços de orquestração permitem que as organizações implantem, monitorem, evoluam e governem agentes autônomos e sistemas agentes em grande escala. As plataformas unem as estruturas fundamentais aos requisitos operacionais do mundo real.

A escolha dos protocolos do agente representa uma decisão estratégica que equilibra as necessidades imediatas de desenvolvimento com flexibilidade e interoperabilidade de longo prazo. Ao priorizar protocolos abertos e criar camadas de abstração apropriadas, as organizações podem criar sistemas de agentes que permaneçam adaptáveis às tecnologias em evolução e, ao mesmo tempo, atendam aos requisitos comerciais atuais.

Para a maioria das organizações, o MCP representa uma base sólida devido ao seu padrão aberto, ecossistema em crescimento, suporte a padrões de agent-to-agent comunicação e recursos de integração de ferramentas. AWS adotou o MCP e o Agent2Agent (A2A) como protocolos estratégicos, contribuindo ativamente para seu desenvolvimento e implementando-os em serviços como o SDK. Strands Agents Ao usar MCP ou A2A junto com ferramentas e meta-ferramentas nativas da estrutura apropriadas, você pode criar sistemas de agentes que ofereçam valor imediato e, ao mesmo tempo, permaneçam adaptáveis às inovações futuras.

Recursos

Use os seguintes AWS e outros recursos relacionados ao desenvolvimento de agentes autônomos.

AWS Blogs

- Amazon Bedrock AgentCore Memory: criando agentes sensíveis ao contexto
- Best practices for building robust generative AI applications with Amazon Bedrock Agents – Part 1
- Best practices for building robust generative AI applications with Amazon Bedrock Agents – Part 2
- Crie pipelines RAG poderosos com o LlamaIndex Amazon Bedrock
- Crie agentes de IA confiáveis com o Amazon Bedrock Observability AgentCore
- Avalie as respostas do RAG com o Amazon Bedrock e o RAGAS LlamaIndex
- Apresentando o Amazon Bedrock AgentCore Code Interpreter
- Apresentando o Amazon Bedrock AgentCore Gateway: transformando o desenvolvimento de ferramentas de agentes de IA corporativos
- Apresentando o Amazon Bedrock AgentCore Identity: protegendo a IA agente em grande escala
- Apresentando Strands Agents um SDK de agentes de IA de código aberto
- Protocolos abertos para interoperabilidade de agentes, parte 1: Comunicação entre agentes no MCP
- Lance e escale com segurança seus agentes e ferramentas no Amazon Bedrock Runtime AgentCore
- AWS Transform para o.NET, o primeiro serviço de IA agente para modernizar aplicativos.NET em grande escala
- AWS Resumo semanal: Strands Agents

AWS Orientação prescritiva

- Operacionalizando a IA agente em AWS
- Fundamentos da IA agêntica em AWS
- Padrões e fluxos de trabalho de IA agentes em AWS
- Construindo arquiteturas sem servidor para IA agente em AWS

- Criação de arquiteturas multilocatárias para IA agente em AWS
- Segurança para IA agente em AWS
- Opções e arquiteturas de geração aumentada de recuperação em AWS

AWS recursos

- Documentação do Amazon Bedrock
- Documentação do Amazon Bedrock AgentCore
- Kit de ferramentas Amazon Bedrock AgentCore Starter (repositório) GitHub
- Documentação do Amazon Nova
- AWS Servidores MCP (GitHubrepositório)

Outros recursos da

- AutoGendocumentação (Microsoft)
- Construindo agentes eficazes (Anthropic)
- CrewAI GitHubrepositório
- Documentação da LangChain
- LangGraphplataforma
- Documentação da LlamaIndex
- Documentação do Model Context Protocol
- Documentação da Strands Agents
- Strands AgentsVisão geral das ferramentas
- Strands AgentsGuia de início rápido

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
Nova seção	Seção de plataformas adicionada	16 de janeiro de 2026
Publicação inicial	—	14 de julho 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

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

integração contínua e entrega contínua (CI/CD)

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

CV

Veja visão computacional.

D

dados em repouso

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

classificação de dados

Um processo para identificar e categorizar os dados em sua rede com base em criticalidade e confidencialidade. É um componente crítico de qualquer estratégia de gerenciamento de riscos de

segurança cibernética, pois ajuda a determinar os controles adequados de proteção e retenção para os dados. A classificação de dados é um componente do pilar de segurança 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 polítca 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. Para obter mais informações, consulte o guia Como quantificar a incerteza em sistemas de aprendizado profundo.

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.