Guia de implementação

Criador de aplicações de IA generativa na AWS

Criador de aplicações de IA generativa na AWS: Guia de implementação

Table of Contents

Essa solução facilita o desenvolvimento, a rápida experimentação e a implantação de aplicativos generativos de inteligência artificial (IA)

O Generative AI Application Builder na AWS facilita o desenvolvimento, a rápida experimentação e a implantação de aplicativos de inteligência artificial (IA) generativa sem exigir uma profunda experiência em IA. Essa solução da AWS acelera o desenvolvimento e simplifica a experimentação, ajudando você a:

- Ingira dados e documentos específicos da sua empresa
- Avalie e compare o desempenho de grandes modelos de linguagem (LLMs)
- Execute tarefas e fluxos de trabalho em várias etapas com agentes de IA
- Crie rapidamente aplicativos extensíveis e implante esses aplicativos com uma arquitetura de nível corporativo

O Generative AI Application Builder na AWS inclui integrações com:

- LLMs disponível na Amazon Bedrock
- LLMs que você implantou na Amazon AI SageMaker
- Bases de conhecimento Amazon Bedrock para geração aumentada de recuperação (RAG)
- Amazon Bedrock Guardrails para implementar proteções e reduzir alucinações
- Amazon Bedrock Agents para criar fluxos de trabalho agentes que podem realizar orquestrações e conclusão de tarefas
- Amazon Bedrock AgentCore para criar, implantar e gerenciar agentes de IA prontos para produção com suporte de tempo de execução estendido
- Servidores Model Context Protocol (MCP) para integração de dados e ferramentas corporativas

Além disso, essa solução permite conexões com o modelo de sua escolha usando LangChain conectores. Esses conectores estão disponíveis em uma função do AWS Lambda que é implantada com a solução. Você pode começar com o assistente de implantação sem código para criar aplicativos generativos de IA para pesquisa conversacional, chatbots gerados por IA, geração de texto e resumo de texto.

Este guia de implementação fornece uma visão geral da solução Generative AI Application Builder na AWS, sua arquitetura e componentes de referência, considerações para planejar a implantação e etapas de configuração para implantar a solução na nuvem da Amazon Web Services (AWS).

Este guia é destinado a arquitetos de soluções, tomadores de decisão de negócios, DevOps engenheiros, cientistas de dados e profissionais de nuvem que desejam implementar o Generative AI Application Builder na AWS em seu ambiente.

Use esta tabela de navegação para encontrar rapidamente respostas para estas perguntas:

Se você deseja...	Leia…
Conhecer o custo da execução dessa solução. O custo estimado para executar essa solução varia com base nos componentes que você implanta e no número de consultas. O custo para executar o painel de implantação com parâmetros padrão e 100 usuários ativos na região Leste dos EUA (Norte da Virgínia) por um mês é de aproximadamente USD 20,12 por mês. O custo de um caso de uso de texto implantado sem RAG para 1 usuário corporativo executando 100 consultas por dia com o LLM é de aproximadamente USD 12,39 por mês. O custo de um caso de uso habilitado para RAG com um índice Amazon Kendra suportando 8.000 interações por dia é de aproximadamente USD 204,26 por mês, mais o custo da base de conhecimento.	Custos
Entenda as considerações de segurança dessa solução.	Segurança
Saiba como planejar cotas para essa solução.	Cotas

Se você deseja...	Leia…
Saiba quais regiões da AWS oferecem suporte a essa solução.	Regiões da AWS com suporte
Visualize ou baixe o CloudFormation modelo da AWS incluído nesta solução para implantar automaticamente os recursos de infraestrutura (a “pilha”) dessa solução.	CloudFormation Modelo da AWS
Acessar o código-fonte e, opcionalmente, usar o AWS Cloud Development Kit (AWS CDK) para implantar a solução.	GitHub repositório

Recursos e benefícios

A solução Generative AI Application Builder na AWS fornece os seguintes recursos:

Experimentação rápida

Essa solução permite que os usuários experimentem rapidamente, eliminando o trabalho pesado necessário para implantar várias instâncias com configurações diferentes e comparar resultados e desempenho. Experimente várias configurações de várias LLMs bases de conhecimento corporativas, de engenharia rápida, grades de proteção, agentes de IA e outros parâmetros.

Escolha e configurabilidade

Com conectores pré-construídos para uma variedade de modelos LLMs, como os disponíveis no Amazon Bedrock, essa solução oferece a flexibilidade de implantar o modelo de sua escolha, bem como a AWS e os principais serviços FM de sua preferência. Você também pode permitir que os Amazon Bedrock Agents cumpram várias tarefas e fluxos de trabalho.

Construtor de agentes

Crie e implante agentes de IA prontos para produção com gerenciamento completo do ciclo de vida. Configure os prompts do sistema, integre os servidores do Model Context Protocol (MCP) para ferramentas corporativas e acesso a dados e habilite recursos de memória para retenção de contexto em conversas. Os agentes são implantados no Amazon Bedrock AgentCore com suporte estendido em tempo de execução e respostas de streaming em tempo real.

Construtor de fluxo

Organize vários agentes do Agent Builder em fluxos de trabalho complexos usando a delegação hierárquica. Crie um agente supervisor que selecione e coordene de forma autônoma agentes especializados do Agent Builder para lidar com tarefas de várias etapas. Configure as descrições dos agentes, as estratégias de delegação e a memória em nível de fluxo de trabalho enquanto reutiliza as implantações existentes do Agent Builder.

Pronto para produção

Criada com os princípios de design do AWS Well-Architected, essa solução oferece segurança e escalabilidade de nível empresarial com alta disponibilidade e baixa latência, garantindo uma integração perfeita em seus aplicativos com altos padrões de desempenho.

Arquitetura modular extensível

Estenda a funcionalidade dessa solução integrando seus projetos existentes ou conectando nativamente outros serviços da AWS. Como esse é um aplicativo de código aberto, você pode usar a camada de LangChain orquestração incluída ou as funções Lambda para se conectar aos serviços de sua escolha.

Integração com o Service Catalog AppRegistry and Application Manager, um recurso do AWS Systems Manager

Essa solução inclui um AppRegistry recurso do Service Catalog para registrar o CloudFormation modelo da solução e seus recursos subjacentes como um aplicativo no AWS Service Catalog AppRegistry e no AWS Systems Manager Application Manager. Com essa integração, você pode gerenciar centralmente os recursos da solução.

Caso de uso do Agent Builder versus Bedrock Agent

Essa solução fornece duas abordagens distintas para trabalhar com agentes de IA, cada uma adequada para diferentes casos de uso e requisitos:

Recurso	Caso de uso do Bedrock Agent	Construtor de agentes
Finalidade	Invoque agentes Amazon Bedrock pré-implantados	Crie, implante e gerencie agentes personalizados

Recurso	Caso de uso do Bedrock Agent	Construtor de agentes
Configuração	Somente ID do agente e ID do alias	Configuração completa do agente: solicitações do sistema, modelos, servidores MCP, memória
Implantação	Camada de invocação simples	Ciclo de vida completo do agente no Runtime AgentCore
Runtime	Serviço Amazon Bedrock Agents	Amazon Bedrock AgentCore com SDK Strands
Integração de ferramentas	Configurado no console do Bedrock Agents	Servidores Model Context Protocol (MCP) e ferramentas Strands integradas
Memória	Gerenciado pela Bedrock Agents (até 30 dias)	AgentCore Memória com retenção configurável de curto e longo prazo
Personalização	Limitado às configurações pré-implantadas do agente	Controle total sobre solicitações, modelos, ferramentas e comportamento
Melhor para	Implantação rápida de agentes existentes	Implantações personalizadas de desenvolvimento e produção de agentes

> **Note**
>
> Ambas as opções oferecem suporte a streaming em tempo real, histórico de conversas e segurança de nível empresarial.

Construtor de fluxo

O Workflow Builder permite a orquestração de vários agentes criando um agente supervisor que delega o trabalho a agentes especializados do Agent Builder. Cada fluxo de trabalho consiste em:

- Agente supervisor: o agente de ponto de entrada que recebe solicitações de usuários e coordena agentes especializados
- Agentes especializados: casos de uso do Agent Builder aos quais o supervisor pode delegar tarefas
- Agentes como padrão de ferramentas: o supervisor registra cada agente do Agent Builder como uma ferramenta e seleciona de forma autônoma quais agentes usar

Recurso	Construtor de agentes	Construtor de fluxo
Finalidade	Crie e implante agentes personali zados únicos	Organize vários agentes do Agent Builder
Tipo de agente	Agente único com ferramentas MCP	Agente supervisor + vários agentes do Agent Builder
Integração de ferramentas	Servidores MCP e ferramentas Strands	Agentes do Agent Builder registrados como ferramentas
Delegação	Invocação direta da ferramenta	Seleção e delegação de agentes autônomos
Complexidade	Tarefas de agente único	Fluxos de trabalho de várias etapas e vários agentes
Reutilização de agentes	N/D	Reutiliza implantações existentes do Agent Builder
Melhor para	Tarefas focadas em um único domínio	Fluxos de trabalho complexos que exigem várias especializações

Note

- Os fluxos de trabalho exigem pelo menos um caso de uso do Agent Builder como agente especializado

- Todos os agentes especializados devem ser casos de uso do Agent Builder implantados no GAAB

Casos de uso

Resposta de perguntas sobre dados corporativos

LLMs e outros modelos básicos foram pré-treinados em um grande corpus de dados, permitindo que tenham um bom desempenho em muitas tarefas de processamento de linguagem natural (PNL). Porém, a maioria dos modelos básicos LLMs são estáticos e foram pré-treinados, limitando sua capacidade de responder com precisão a perguntas sobre tópicos novos, especializados ou proprietários. Usando o aprendizado baseado em prompts, você pode aproveitar os poderosos recursos de NLP e geração de texto de um LLM para fornecer experiências mais ricas aos clientes em relação aos dados da sua empresa.

Prototipagem generativa rápida de IA

Pronta para uso, a solução vem com vários fornecedores de modelos e casos de uso. Com um assistente de implantação fácil de usar, os clientes podem implantar casos de uso pré-criados para permitir a rápida experimentação de diferentes protótipos e cargas de trabalho generativas de IA.

Comparação e experimentação do Multi LLM

LLMs tenha um desempenho diferente e, dadas as necessidades específicas do seu aplicativo, você pode descobrir que um LLM se adapta melhor ao seu aplicativo do que outro. Isso pode ser por motivos relacionados ao desempenho, precisão, custo, criatividade ou muitos outros fatores. Essa solução permite que você implante rapidamente vários casos de uso, permitindo que você experimente e compare diferentes configurações até encontrar o que atende às suas necessidades.

Conceitos e definições

Esta seção descreve os conceitos básicos e define a terminologia específica desta solução:

usuário administrador

No contexto deste guia, o usuário administrador é o responsável por gerenciar o conteúdo contido na implantação. Esse usuário obtém acesso à interface do usuário do painel de implantação e é

o principal responsável por organizar a experiência do usuário comercial. Este é nosso principal cliente-alvo.

usuário comercial

No contexto deste guia, o usuário corporativo representa as pessoas para as quais o caso de uso foi implantado. Eles são os consumidores da base de conhecimento e os clientes responsáveis por avaliar e experimentar o. LLMs

Painel de implantação

O painel de implantação é uma interface web que serve como um console de gerenciamento para usuários administradores visualizarem, gerenciarem e criarem seus casos de uso. Esse painel permite que os clientes experimentem, iterem e produzam rapidamente várias AI/ML cargas de trabalho aproveitando. LLMs

DevOps usuário

No contexto deste guia, o DevOps usuário é o único responsável por implantar a solução na conta da AWS e por gerenciar a infraestrutura, atualizar a solução, monitorar o desempenho e manter a integridade geral e o ciclo de vida da solução.

caso de uso

Os casos de uso são aplicativos isolados da solução geral que se integram LLMs para permitir experiências mais ricas ao cliente, permitindo a adição de uma interface de linguagem natural em aplicativos novos ou existentes. Os casos de uso podem ser implantados por meio do painel de implantação ou por conta própria.

> **Note**
>
> Para obter uma referência geral dos termos da AWS, consulte o Glossário da AWS.

Visão geral da arquitetura

Esta seção fornece diagramas de arquitetura de implementação de referência para os componentes implantados com essa solução.

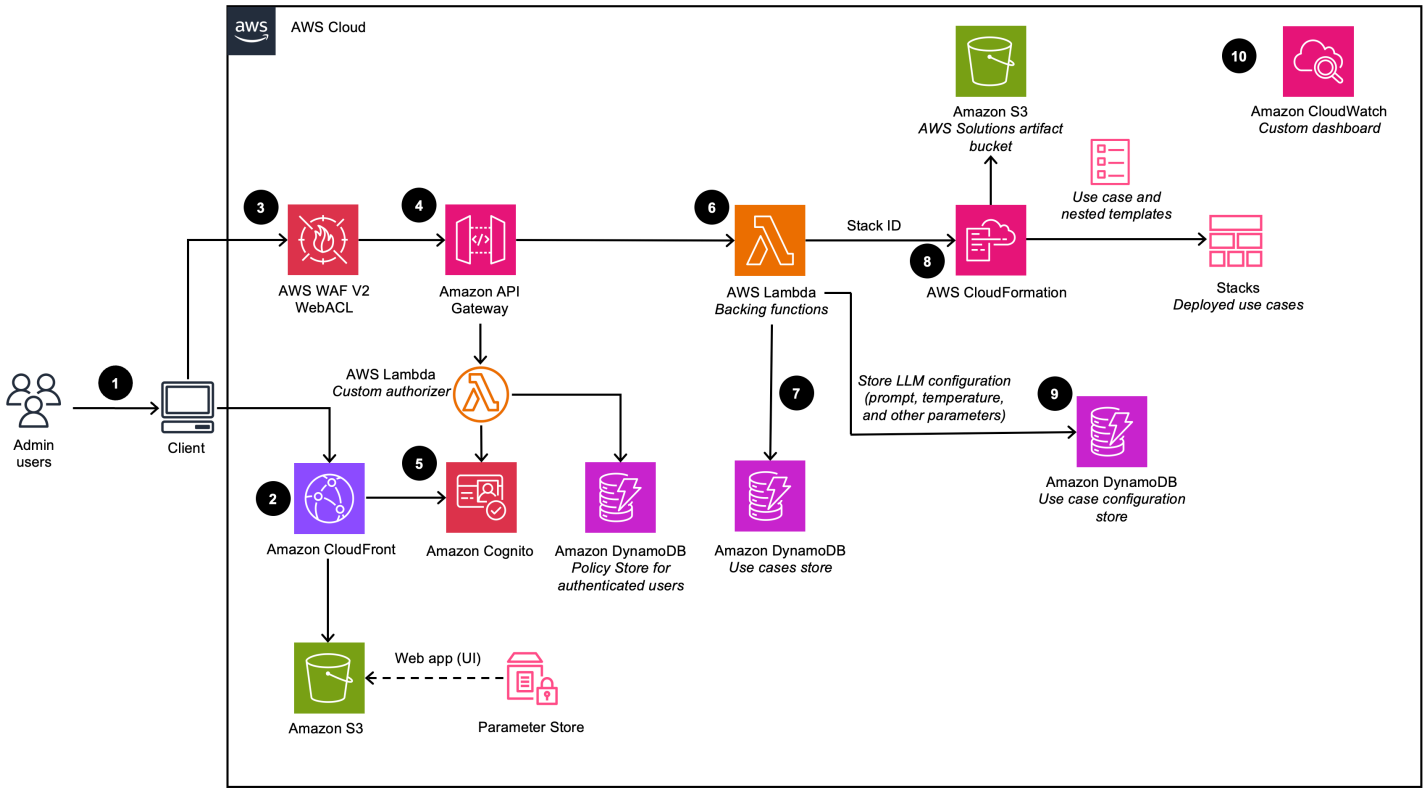
Diagramas de arquitetura

Para oferecer suporte a vários casos de uso e necessidades comerciais, essa solução fornece seis CloudFormation modelos da AWS:

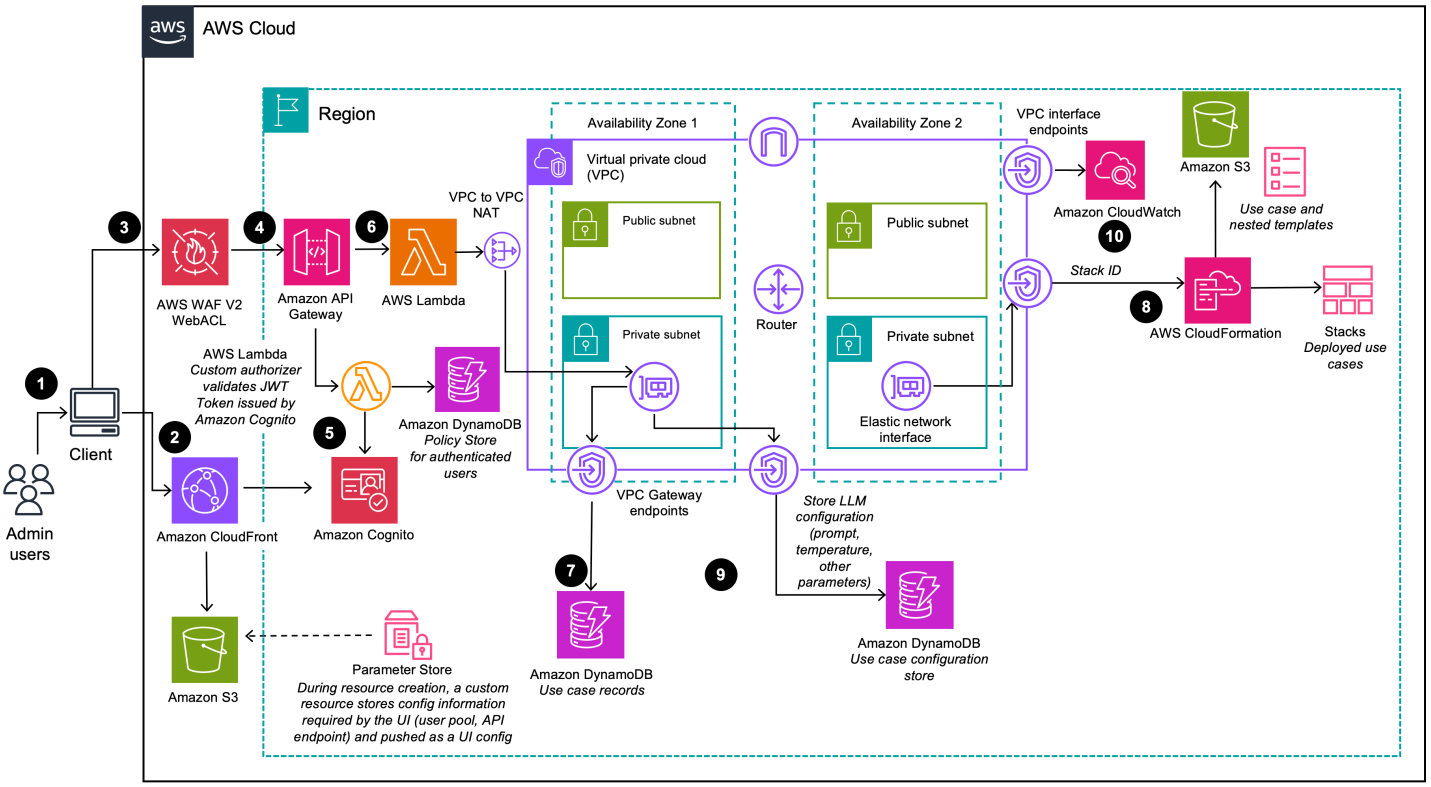
1. Painel de implantação - O painel de implantação é uma interface web que serve como um console de gerenciamento para usuários administradores visualizarem, gerenciarem e criarem seus casos de uso. Esse painel permite que os clientes experimentem, iterem e produzam rapidamente várias AI/ML cargas de trabalho aproveitando. LLMs
2. Caso de uso de texto - O caso de uso de texto permite que os usuários experimentem uma interface de linguagem natural usando IA generativa. Esse caso de uso pode ser integrado a aplicativos novos ou existentes e pode ser implantado por meio do painel de implantação ou de forma independente por meio de uma URL fornecida.
3. Caso de uso do Bedrock Agent - O caso de uso do Bedrock Agent permite o uso dos Bedrock Agents existentes para concluir tarefas ou automatizar fluxos de trabalho repetidos.
4. Servidor MCP - O caso de uso do MCP Server permite a implantação e o gerenciamento de servidores do Model Context Protocol que fornecem ferramentas padronizadas e acesso a recursos para aplicativos de IA. Suporta métodos de gateway para agrupar funções APIs Lambda existentes e servidores MCP externos e métodos de tempo de execução para implantar servidores MCP em contêineres personalizados.
5. Agent Builder — O Agent Builder permite a criação e a implantação de agentes de IA prontos para produção no Amazon Bedrock AgentCore com controle total de configuração, integração de servidores MCP e recursos de gerenciamento de memória.
6. Construtor de fluxo de trabalho - O criador de fluxo de trabalho permite a criação de agentes supervisores que orquestram vários agentes do Agent Builder usando o padrão de delegação de Agentes como Ferramentas para fluxos de trabalho complexos com vários agentes.

Painel de implantação

Descreve a arquitetura do painel de implantação (quando implantado com a opção VPC desativada)



Descreve a arquitetura do painel de implantação (quando implantado com a opção VPC ativada)



> **Note**
>
> Os CloudFormation recursos da AWS são criados a partir de construções do AWS Cloud Development Kit (AWS CDK).

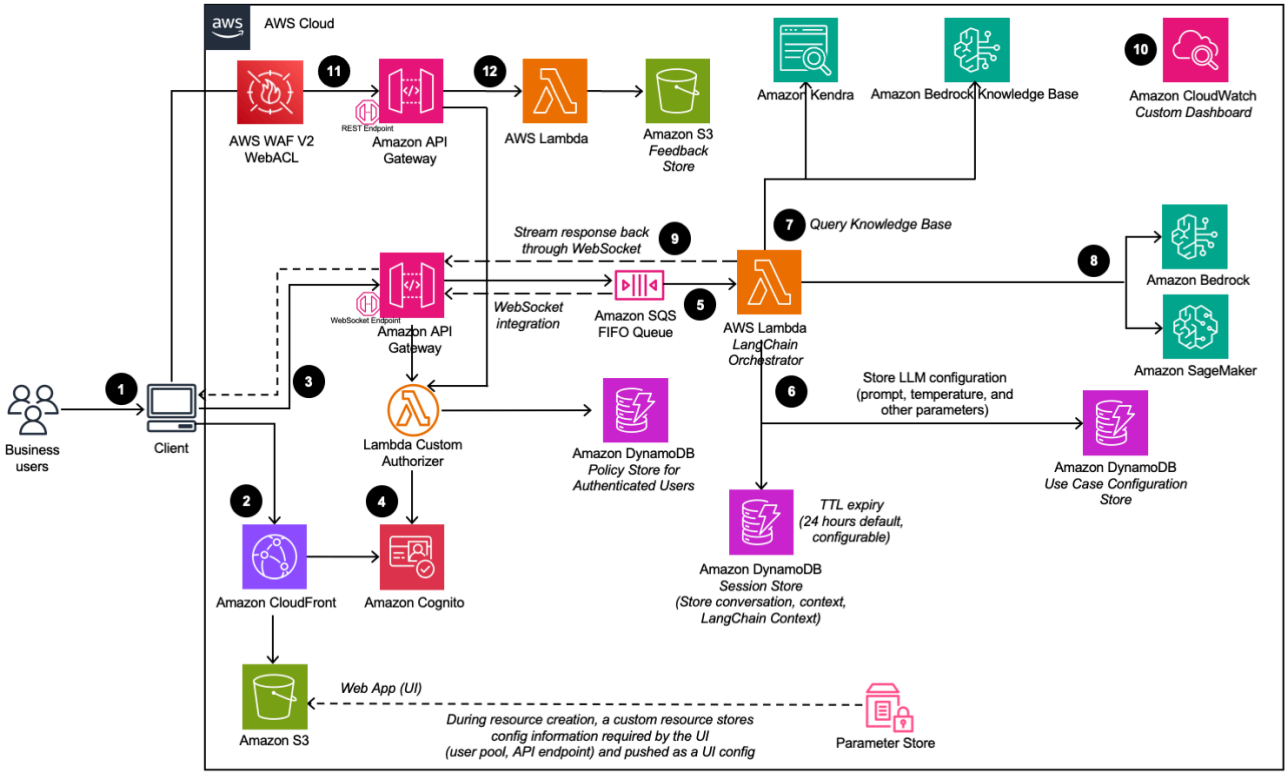
O fluxo de processo de alto nível para os componentes da solução implantados com o CloudFormation modelo da AWS é o seguinte:

1. Os usuários administradores fazem login na interface de usuário (UI) do Deployment Dashboard.
2. A Amazon CloudFront fornece a interface web, que é hospedada em um bucket do Amazon Simple Storage Service (Amazon S3).
3. O AWS WAF os APIs protege contra ataques. Essa solução configura um conjunto de regras chamado lista de controle de acesso à web (Web ACL) que permite, bloqueia ou conta solicitações da web com base em regras e condições de segurança da web configuráveis e definidas pelo usuário.
4. A interface web utiliza um conjunto de REST APIs que são expostos usando o Amazon API Gateway.
5. O Amazon Cognito autentica usuários e faz backup tanto CloudFront da interface de usuário da web quanto do API Gateway.
6. O AWS Lambda fornece a lógica de negócios para os endpoints REST. Essa função de apoio do Lambda gerencia e cria os recursos necessários para realizar implantações de casos de uso usando a AWS. CloudFormation
7. O Amazon DynamoDB armazena a lista de implantações.
8. Quando um novo caso de uso é criado pelo usuário administrador, a função Lambda de apoio inicia CloudFormation um evento de criação de pilha para o caso de uso solicitado.
9. Todas as opções de configuração do LLM fornecidas pelo usuário administrador no assistente de implantação são salvas no DynamoDB. A implantação usa essa tabela do DynamoDB para configurar o LLM em tempo de execução.
10. Usando a Amazon CloudWatch, essa solução coleta métricas operacionais de vários serviços para gerar painéis personalizados que permitem monitorar o desempenho e a integridade operacional da solução.

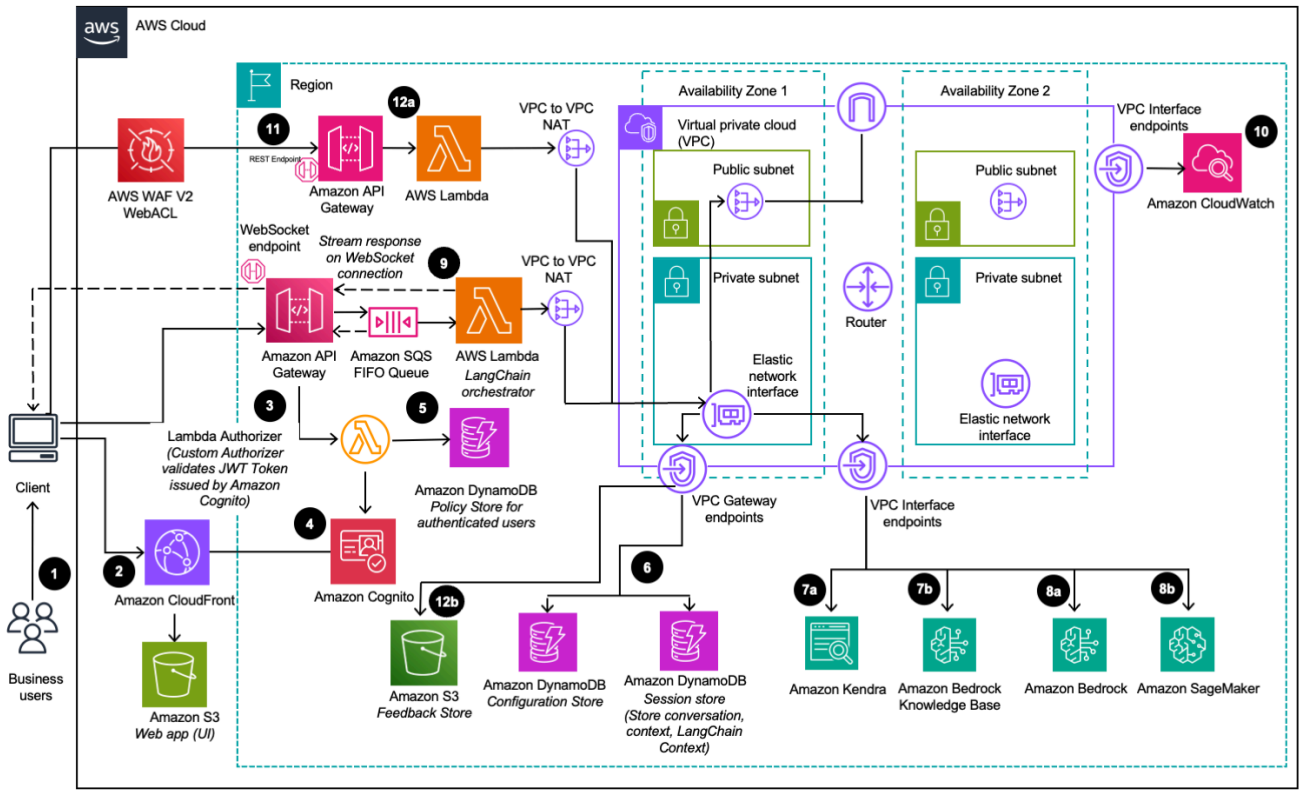
> **Note**
>
> - Se você optar por implantar essa solução em uma Amazon VPC, os dados serão roteados dentro da sua rede privada.
> - Embora o painel de implantação possa ser lançado na maioria das regiões da AWS, os casos de uso implantados têm certas restrições com base na disponibilidade do serviço. Consulte Regiões compatíveis da AWS para obter mais detalhes.

Caso de uso de texto

Representa a arquitetura do caso de uso do Text (quando implantada com a opção VPC desativada)



Representa a arquitetura do caso de uso do Text (quando implantada com a opção VPC ativada)



O fluxo de processo de alto nível para os componentes da solução implantados com o CloudFormation modelo da AWS é o seguinte:

1. Os usuários administradores implantam o caso de uso usando o Painel de Implantação. Os usuários corporativos fazem login na interface do usuário do caso de uso.
2. CloudFront fornece a interface do usuário da web que está hospedada em um bucket S3.
3. A interface de usuário da web aproveita uma WebSocket integração criada usando o API Gateway. O API Gateway é apoiado por uma função autorizadora personalizada do Lambda, que retorna a política apropriada do AWS Identity and Access Management (IAM) com base no grupo Amazon Cognito ao qual o usuário autenticador pertence. A política é armazenada no DynamoDB.
4. O Amazon Cognito autentica usuários e faz backup tanto CloudFront da interface de usuário da web quanto do API Gateway.
5. As solicitações recebidas do usuário corporativo são passadas do API Gateway para uma fila do Amazon SQS e, em seguida, para o orquestrador. LangChain O LangChain orquestrador é uma coleção de funções e camadas do Lambda que fornecem a lógica de negócios para atender às solicitações provenientes do usuário corporativo. A fila permite a operação assíncrona da integração do API Gateway com o Lambda. A fila passa as informações de conexão para as

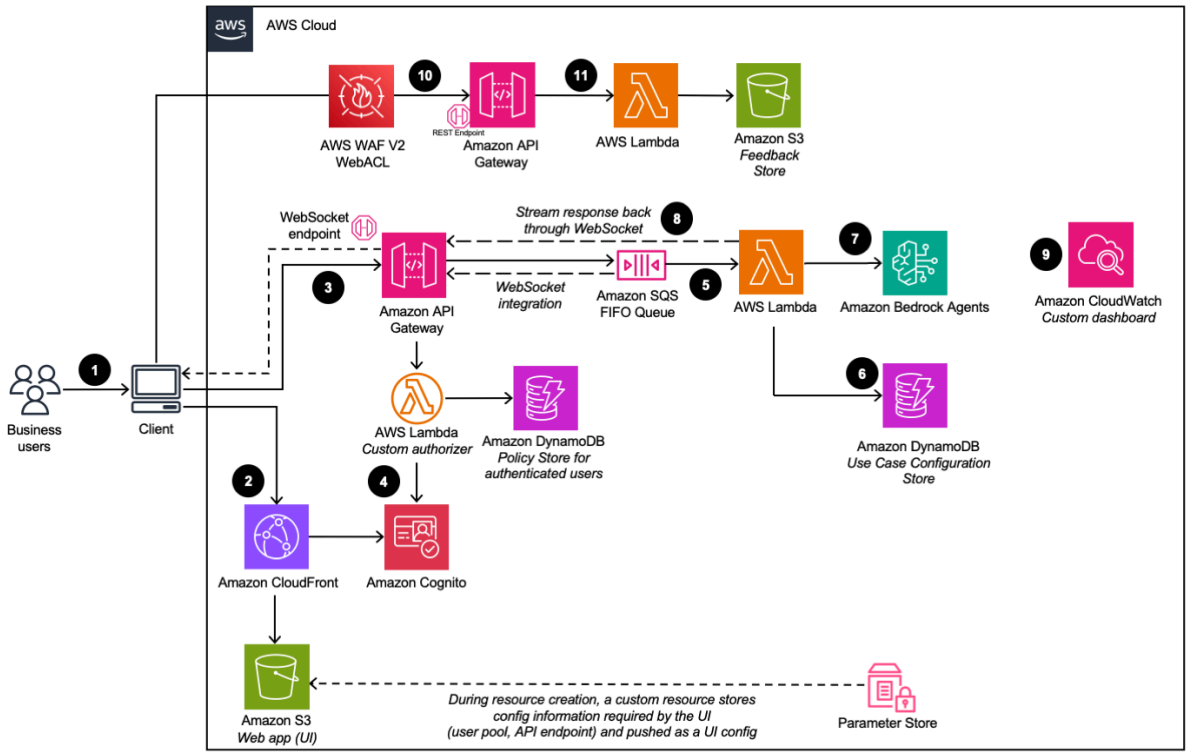
funções do Lambda, que então publicam os resultados diretamente na conexão websocket do API Gateway para suportar chamadas de inferência de longa duração.

6. O LangChain orquestrador usa o Amazon DynamoDB para obter as opções configuradas do LLM e as informações necessárias da sessão (como o histórico do bate-papo).
7. Se a implantação tiver uma base de conhecimento habilitada, o LangChain orquestrador aproveita o Amazon Kendra ou as bases de conhecimento do Amazon Bedrock para executar uma consulta de pesquisa para recuperar trechos de documentos.
8. Usando o histórico de bate-papo, a consulta e o contexto da base de conhecimento, o LangChain orquestrador cria a solicitação final e envia a solicitação para o LLM hospedado no Amazon Bedrock ou no Amazon AI. SageMaker
9. Quando a resposta volta do LLM, o LangChain orquestrador transmite a resposta de volta pelo API Gateway WebSocket para ser consumida pelo aplicativo cliente.
10. Usando a Amazon CloudWatch, essa solução coleta métricas operacionais de vários serviços para gerar painéis personalizados que permitem monitorar o desempenho e a integridade operacional da implantação.
11. Se a coleta de feedback estiver ativada, um endpoint da API REST, utilizando o Amazon API Gateway, será disponibilizado para a coleta de feedback do usuário.
12. O feedback de apoio lambda aumenta o feedback enviado com metadados adicionais específicos do caso de uso (por exemplo, modelo usado) e armazena os dados no Amazon S3 para análise e geração de relatórios posteriores pelos usuários. DevOps

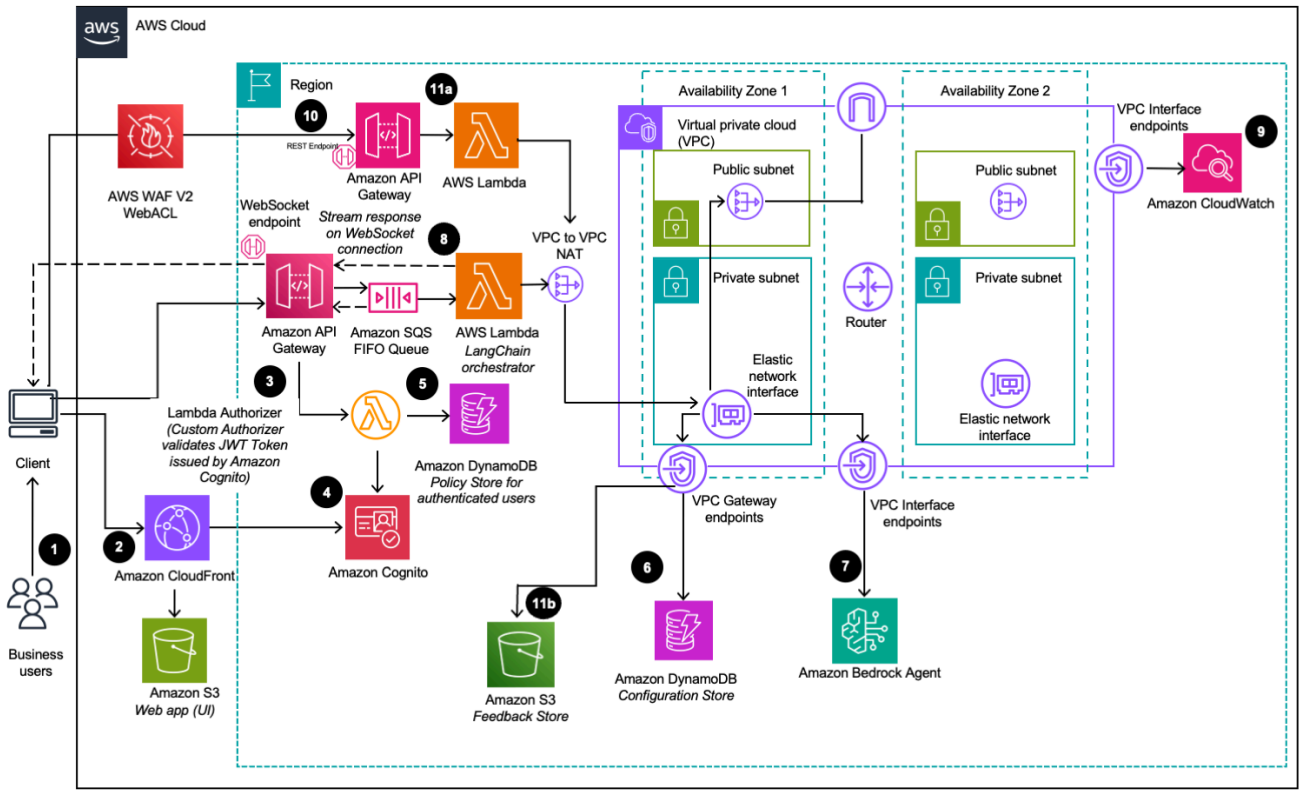
> **Note**
>
> Se você optar por implantar essa solução em uma Amazon VPC, os dados serão roteados para sua rede privada.

Caso de uso do Bedrock Agent

Descreve a arquitetura do caso de uso do Bedrock Agent (quando implantado com a opção VPC desativada)



Descreve a arquitetura do caso de uso do Bedrock Agent (quando implantado com a opção VPC ativada)



O fluxo de processo de alto nível para os componentes da solução implantados com o CloudFormation modelo da AWS é o seguinte:

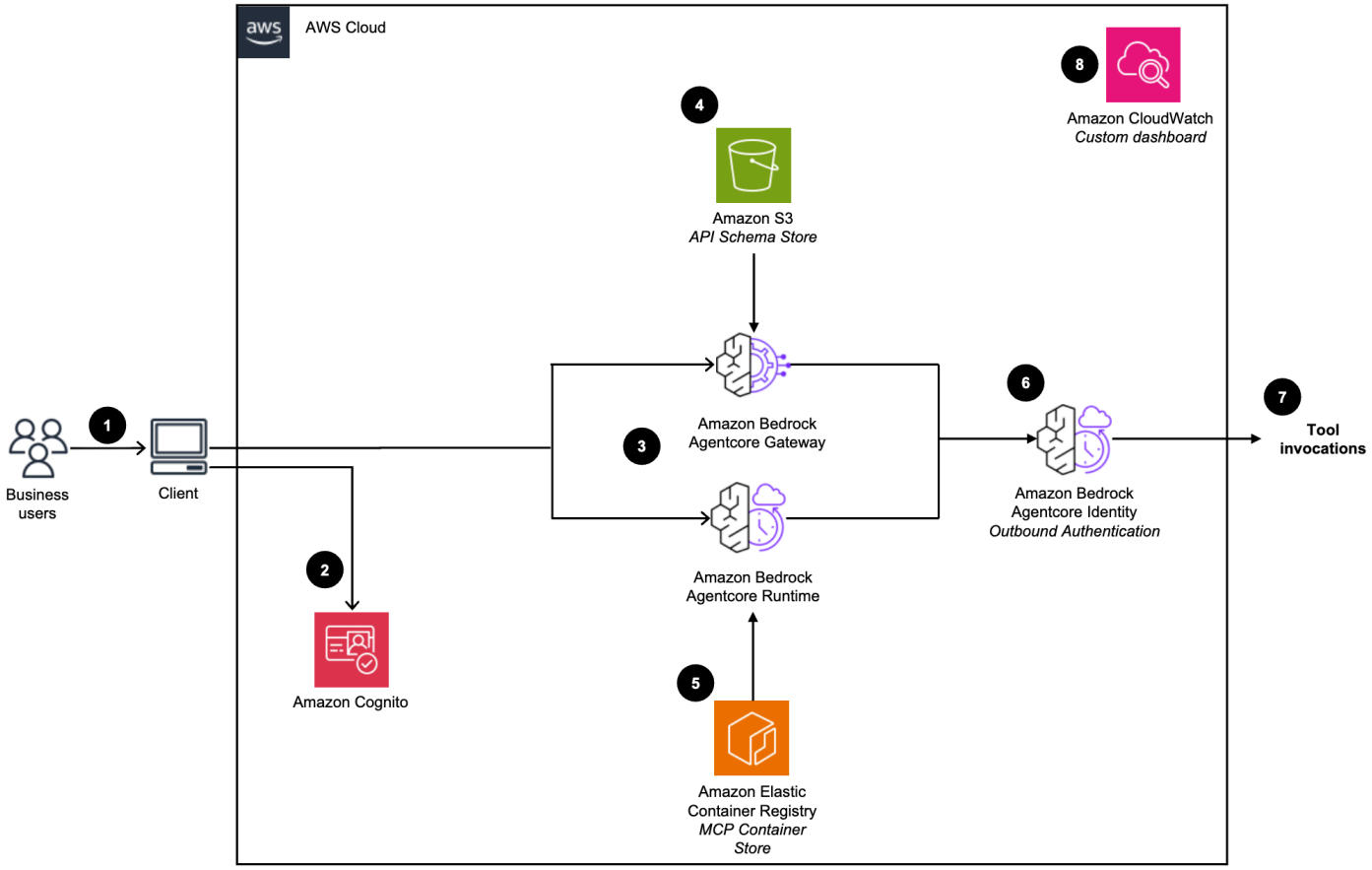
1. Os usuários administradores implantam o caso de uso usando o Painel de Implantação. Os usuários corporativos fazem login na interface do usuário do caso de uso.
2. CloudFront fornece a interface do usuário da web que está hospedada em um bucket S3.
3. A interface de usuário da web aproveita uma WebSocket integração criada usando o API Gateway. O API Gateway é apoiado por uma função autorizadora personalizada do Lambda, que retorna a política apropriada do AWS Identity and Access Management (IAM) com base no grupo Amazon Cognito ao qual o usuário autenticador pertence. A política é armazenada no DynamoDB.
4. O Amazon Cognito autentica usuários e faz backup tanto CloudFront da interface de usuário da web quanto do API Gateway.
5. As solicitações recebidas do usuário corporativo são passadas do API Gateway para uma fila do Amazon SQS e, em seguida, para a função AWS Lambda. A fila permite a operação assíncrona da integração do API Gateway com o Lambda. A fila passa as informações de conexão para a função Lambda, que então publicará os resultados diretamente na conexão websocket do API Gateway para suportar chamadas de inferência de longa duração.

6. A função AWS Lambda usa o Amazon DynamoDB para obter as configurações do caso de uso conforme necessário.
7. Usando a entrada do usuário e qualquer configuração de caso de uso relevante, a função AWS Lambda cria e envia uma carga de solicitação para o Amazon Bedrock Agent configurado para cumprir a intenção do usuário.
8. Quando a resposta volta do Amazon Bedrock Agent, a função Lambda transmite a resposta de volta por meio do API WebSocket Gateway para ser consumida pelo aplicativo cliente.
9. Usando a Amazon CloudWatch, essa solução coleta métricas operacionais de vários serviços para gerar painéis personalizados que permitem monitorar o desempenho e a integridade operacional da implantação.
10. Se a coleta de feedback estiver ativada, um endpoint da API REST, utilizando o Amazon API Gateway, será disponibilizado para a coleta de feedback do usuário.
11. O feedback de apoio lambda aumenta o feedback enviado com metadados adicionais específicos do caso de uso e armazena os dados no Amazon S3 para análise e geração de relatórios posteriores pelos usuários. DevOps

> **Note**
>
> Se você optar por implantar essa solução em uma Amazon VPC, os dados serão roteados dentro da sua rede privada.

Caso de uso do MCP Server

Descreve a arquitetura do caso de uso do MCP Server



O caso de uso do MCP Server permite a implantação e o gerenciamento de servidores do Model Context Protocol no Amazon Bedrock AgentCore. Os servidores MCP fornecem uma interface padronizada para aplicativos de IA acessarem ferramentas, recursos e fontes de dados corporativos.

A solução oferece suporte a dois métodos de implantação:

- Método de gateway: agrupa funções Lambda existentes, APIs REST ou servidores MCP externos como ferramentas MCP, manipulando a tradução de protocolos automaticamente
- Método de tempo de execução: implanta servidores MCP personalizados em contêineres a partir de imagens do Amazon ECR

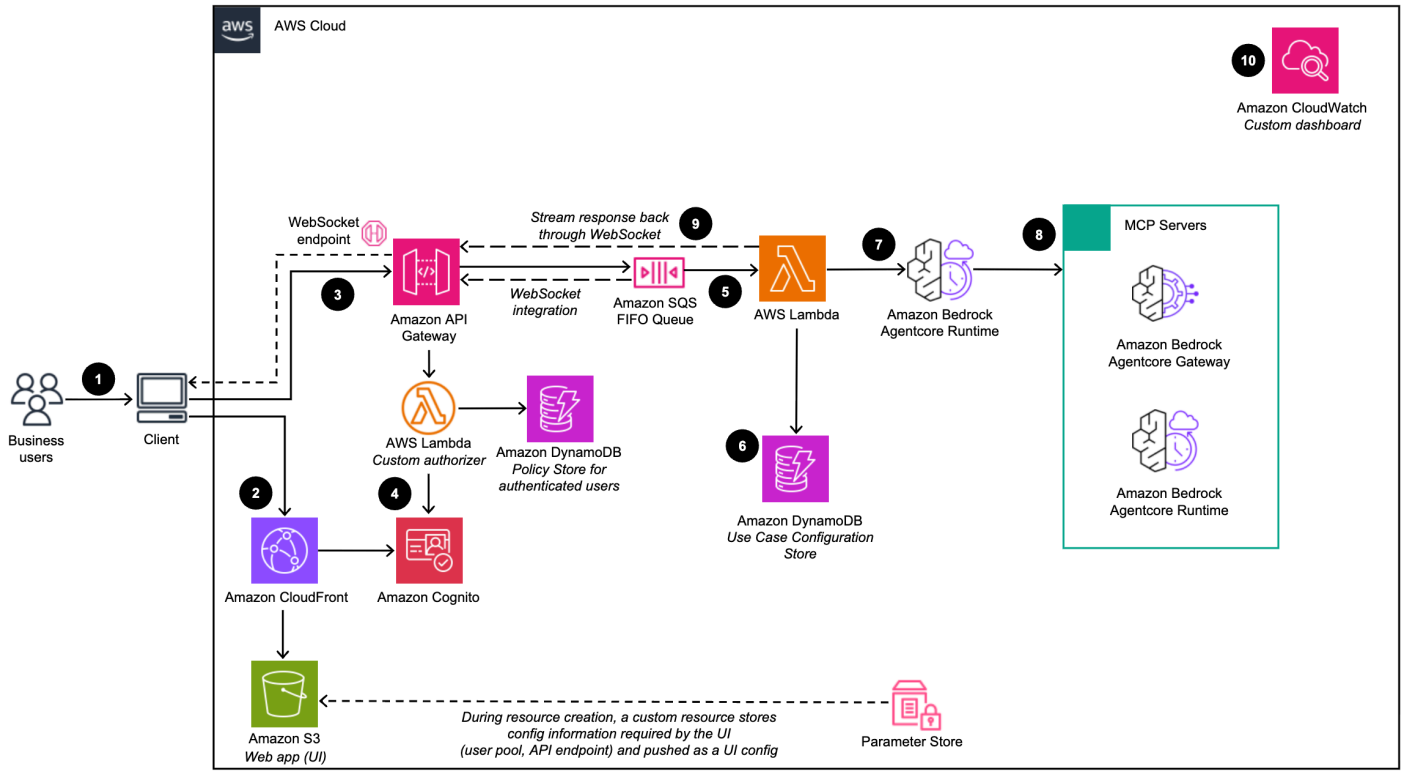
O fluxo de processo de alto nível para a implantação do MCP Server é o seguinte:

1. Os usuários administradores implantam o caso de uso do MCP Server usando o Deployment Dashboard, selecionando o método de implantação Gateway ou Runtime.
2. Essa ação é autenticada com o Amazon Cognito.

3. Para a implantação do Gateway, a solução cria um Amazon Bedrock AgentCore Gateway que transforma funções Lambda existentes ou servidores MCP externos em APIs ferramentas compatíveis com MCP. Para a implantação do Runtime, a solução implanta servidores MCP em contêineres no Amazon Bedrock AgentCore Runtime usando imagens ECR fornecidas.
4. As implantações de gateway recuperam os API/Lambda/Smithy esquemas necessários de seu local de upload no Amazon S3 ou se conectam diretamente aos endpoints de URL do servidor MCP.
5. As implantações em tempo de execução recuperam o servidor MCP em contêineres fornecido pelo usuário do Amazon Elastic Container Registry (ECR)
6. O MCP Server é instrumentado com um cliente Amazon Bedrock Identity AgentCore OAuth
7. O MCP Server disponibiliza as ferramentas associadas no endpoint /mcp para que os agentes as descubram.
8. A Amazon CloudWatch coleta métricas e registros operacionais de implantações de servidores MCP para monitoramento e solução de problemas.

Caso de uso do Agent Builder

Representa a arquitetura do Agent Builder



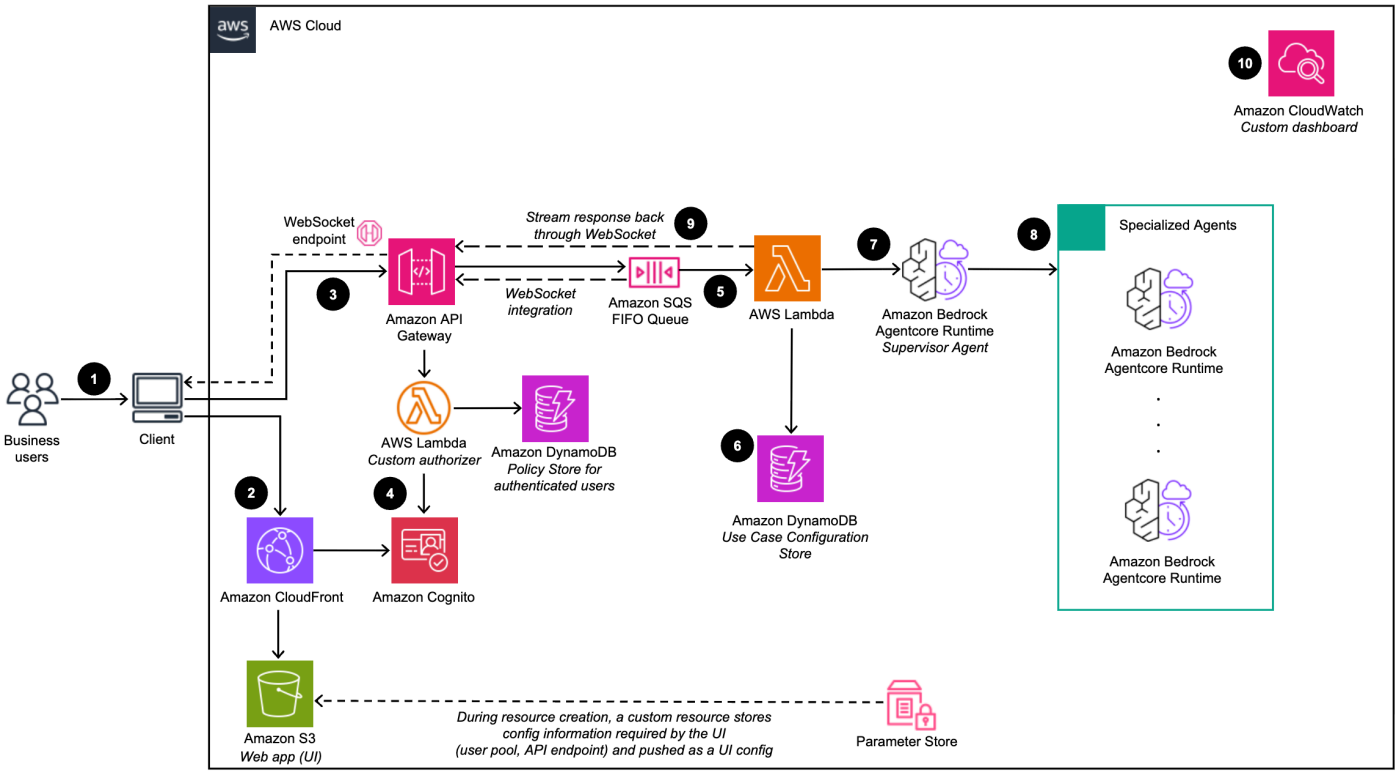
O fluxo de processo de alto nível para os componentes do Agent Builder implantados com o CloudFormation modelo da AWS é o seguinte:

1. Os usuários administradores implantam o caso de uso usando o Painel de Implantação. Os usuários corporativos fazem login na interface do usuário do caso de uso.
2. CloudFront fornece a interface do usuário da web que está hospedada em um bucket S3.
3. A interface de usuário da web aproveita uma WebSocket integração criada usando o API Gateway. O API Gateway é apoiado por uma função autorizadora personalizada do Lambda, que retorna a política apropriada do AWS Identity and Access Management (IAM) com base no grupo Amazon Cognito ao qual o usuário autenticador pertence. A política é armazenada no DynamoDB.
4. O Amazon Cognito autentica usuários e faz backup tanto CloudFront da interface de usuário da web quanto do API Gateway.
5. As solicitações recebidas do usuário corporativo são passadas do API Gateway para uma fila do Amazon SQS e, em seguida, para a função AWS Lambda. A fila permite a operação assíncrona da integração do API Gateway com o Lambda. A fila passa as informações de conexão para a função Lambda, que então publicará os resultados diretamente na conexão websocket do API Gateway para suportar chamadas de inferência de longa duração.
6. A função AWS Lambda recupera a configuração do agente do DynamoDB.
7. Usando a entrada do usuário e qualquer configuração de caso de uso relevante, a função AWS Lambda cria e envia uma carga de solicitação ao agente, executada no Amazon Bedrock Runtime. AgentCore
8. O agente se conecta aos servidores MCP associados e registra as ferramentas na instância do agente de strings. O agente então seleciona e executa ações de forma autônoma com base nas descrições das ferramentas e nos requisitos da tarefa.
9. Quando a resposta volta do tempo de AgentCore execução do Amazon Bedrock, a função Lambda transmite a resposta de volta por meio do API WebSocket Gateway para ser consumida pelo aplicativo cliente.

> **Note**
>
> - O processamento do agente é limitado ao tempo limite de execução do Lambda (15 minutos).

Caso de uso do Workflow Builder

Descreve a arquitetura do Workflow Builder



O fluxo de processo de alto nível para os componentes do Workflow Builder implantados com o CloudFormation modelo da AWS é o seguinte:

1. Os usuários administradores implantam o fluxo de trabalho usando o Painel de Implantação, selecionando agentes do Agent Builder para incluir como agentes especializados.
2. CloudFront fornece a interface do usuário da web que está hospedada em um bucket S3.
3. A interface de usuário da web aproveita uma WebSocket integração criada usando o API Gateway. O API Gateway é apoiado por uma função autorizadora personalizada do Lambda, que retorna a política apropriada do AWS Identity and Access Management (IAM) com base no grupo Amazon Cognito ao qual o usuário autenticador pertence. A política é armazenada no DynamoDB.
4. O Amazon Cognito autentica usuários e faz backup tanto CloudFront da interface de usuário da web quanto do API Gateway.
5. As solicitações recebidas do usuário corporativo são passadas do API Gateway para uma fila do Amazon SQS e, em seguida, para a função AWS Lambda. A fila permite a operação assíncrona da integração do API Gateway com o Lambda.

6. A função AWS Lambda recupera a configuração do fluxo de trabalho do DynamoDB, incluindo a lista de agentes especializados do Agent Builder.
7. Usando a entrada do usuário e a configuração do fluxo de trabalho, o Lambda envia solicitações para o Amazon Bedrock AgentCore Runtime que hospeda o agente supervisor.
8. O agente supervisor cria instâncias locais de todos os agentes especializados do Agent Builder no ambiente AgentCore Runtime. Esses agentes especializados são registrados como ferramentas usando o padrão Agents as Tools. O supervisor então seleciona e delega de forma autônoma o trabalho a agentes especializados com base nas descrições dos agentes e nos requisitos da tarefa.
9. O agente supervisor agrega resultados de agentes especializados e formula a resposta final, devolvendo-a ao Lambda para ser transmitida de volta ao aplicativo cliente por meio do Websocket do API Gateway.

> **Note**
>
> - O processamento do fluxo de trabalho é limitado ao tempo limite de execução do Lambda (15 minutos).

Considerações de design do AWS Well-Architected

Essa solução foi projetada com as melhores práticas do AWS Well-Architected Framework, que ajuda os clientes a projetar e operar cargas de trabalho confiáveis, seguras, eficientes e econômicas na nuvem.

Esta seção descreve como os princípios de design e as melhores práticas do Well-Architected Framework foram aplicados ao criar essa solução.

Excelência operacional

Esta seção descreve como arquitetamos essa solução usando os princípios e as melhores práticas do pilar de excelência operacional.

- Criamos a solução infrastructure-as-code usando a Amazon CloudFormation.
- As funções Lambda enviam métricas personalizadas CloudWatch e um CloudWatch painel personalizado para monitorar a integridade da solução.

- Os componentes da solução são altamente modularizados, oferecendo a flexibilidade de escolher quais componentes implantar.

Segurança

Esta seção descreve como arquitetamos essa solução usando os princípios e as melhores práticas do pilar de segurança.

- O painel de implantação e todos os casos de uso são autenticados e autorizados com o Amazon Cognito.
- Todas as comunicações entre serviços usam funções do AWS IAM.
- Todas as funções da solução seguem o acesso com privilégios mínimos; ou seja, somente as permissões mínimas necessárias são concedidas.
- Todo o armazenamento de dados, incluindo buckets S3, DynamoDB e Amazon Kendra, tem criptografia em repouso.

Confiabilidade

Esta seção descreve como arquitetamos essa solução usando os princípios e as melhores práticas do pilar de confiabilidade.

- Arquitetura baseada no paradigma sem servidor.
- Criamos a arquitetura para escalabilidade horizontal sob demanda e recuperação automática de falhas na infraestrutura subjacente.
- A arquitetura inclui solicitações de buffer e limitação para não sobrecarregar os endpoints subjacentes.

Eficiência de desempenho

Esta seção descreve como arquitetamos essa solução usando os princípios e as melhores práticas do pilar de eficiência de desempenho.

- A solução usa o DynamoDB, um banco de dados NoSQL sem servidor totalmente gerenciado com escalabilidade sob demanda.
- A solução usa o Amazon S3 para armazenamento de objetos e para hospedar um site (por meio de CloudFront) para oferecer baixo custo, escalabilidade e durabilidade de 11 9s.

Otimização de custos

Esta seção descreve como arquitetamos essa solução usando os princípios e as melhores práticas do pilar de otimização de custos.

- Sempre que possível, criamos a solução para usar a arquitetura sem servidor; portanto, você paga apenas pelo que usa.

Sustentabilidade

Esta seção descreve como arquitetamos essa solução usando os princípios e as melhores práticas do pilar de sustentabilidade.

- A arquitetura modular e em componentes da solução oferece a flexibilidade de personalizar recursos a serem provisionados para casos de uso individuais.
- A arquitetura usa computação e armazenamento sem servidor, o que otimiza a utilização dos recursos.
- Como uma solução baseada em nuvem, essa solução se beneficia de recursos compartilhados, redes, energia, resfriamento e instalações físicas.

Detalhes de arquitetura

Esta seção descreve os componentes e os serviços da AWS que compõem essa solução e os detalhes da arquitetura sobre como esses componentes funcionam juntos.

Serviços da AWS nesta solução

Serviço da AWS	Description
Amazon API Gateway	Principal. Esse serviço fornece o REST APIs para o painel de implantação e a WebSocket API para o caso de uso.
AWS CloudFormation	Principal. Essa solução é distribuída como um CloudFormation modelo e CloudFormation implanta os recursos da AWS para a solução.
Amazon CloudFront	Núcleo. CloudFront serve o conteúdo da web hospedado no Amazon S3.
Amazon Cognito	Principal. Esse serviço gerencia o gerenciam ento e a autenticação de usuários para a API.
Amazon DynamoDB	Principal. O DynamoDB armazena informaçõ es de implantação e detalhes de configuração para o painel de implantação. Ele armazena o histórico de conversas e conversas IDs no caso de uso de texto para permitir o histórico de conversas e a desambiguação de consultas.
AWS Lambda	Principal. A solução usa funções Lambda para: * Apoie os endpoints de WebSocket REST e API * Gerencie a lógica central de cada orquestrador de casos de uso * Implemente recursos personalizados durante a implantação CloudFormation

Serviço da AWS	Description
Amazon S3	Principal. O Amazon S3 hospeda o conteúdo estático da web.
Amazon CloudWatch	Suporte. Essa solução publica registros dos recursos da solução no CloudWatch Logs e publica métricas nas métricas. CloudWatch A solução também cria um CloudWatch painel para visualizar esses dados.
AWS Systems Manager	Suporte. O Systems Manager fornece monitoramento de recursos em nível de aplicativo e visualização de operações de recursos e dados de custos. Também usado para armazenar dados de configuração no Parameter Store.
AWS WAF	Suporte. O AWS WAF é implantado na frente da implantação do API Gateway para protegê-l o.
Amazon Bedrock	Opcional. A solução utiliza o Amazon Bedrock para acessar modelos básicos ou personali zados, Amazon Bedrock Agents e Amazon Bedrock Knowledge Bases. O Amazon Bedrock é a integração recomendada para evitar que seus dados saiam da rede AWS.
Amazon Bedrock AgentCore	Opcional A solução utiliza o Amazon Bedrock AgentCore para executar e oferecer suporte às conexões do MCP Server, bem como aos casos de uso do Agent Builder e do Workflow.

Serviço da AWS	Description
Amazon Elastic Container Registry (Amazon ECR)	Opcional. Para implantações do Agent Builder, o ECR armazena e distribui imagens de contêiner do agente. A solução usa o ECR Pull-Through Cache para recuperar automatic amente imagens pré-criadas do agente do repositório público de ECR da equipe do GAAB.
Distro da AWS para OpenTelemetry (ADOT)	Opcional. Para implantações do Agent Builder, o ADOT fornece instrumentação automática para a observabilidade do agente, permitindo rastreamento distribuído e registro estruturado para as operações do agente.
Amazon Kendra	Opcional. No caso de uso do Text, os usuários administradores podem opcionalmente decidir conectar um índice do Amazon Kendra para usar como base de conhecimento para a conversa com o LLM. Isso pode ser usado para injetar novas informações no LLM, dando-lhe a capacidade de usar essas informações em suas respostas.
SageMaker Inteligência Artificial da Amazon	Opcional. A solução pode ser integrada a um endpoint de inferência de SageMaker IA da Amazon para acesso FMs que está hospedado em sua conta e região da AWS e é uma integração preferencial para evitar que seus dados saiam da rede da AWS. **Note** Você deve implantar a solução na mesma região em que o endpoint de inferência está disponível.

Serviço da AWS	Description
Amazon Virtual Private Cloud	Opcional. A solução oferece a opção de implantar componentes com uma configuração habilitada para VPC. Ao implantar a solução com uma configuração habilitada para VPC, você tem a opção de permitir que a solução crie uma VPC para você ou use uma VPC existente que exista na mesma conta e região em que a solução será implantada (traga sua própria VPC). Se a solução criar a VPC, ela criará os componentes de rede necessários que incluem sub-redes, grupos de segurança e suas regras, tabelas de rotas, rede, gateways NAT, gateways de Internet ACLs, endpoints de VPC e suas políticas.

Painel de implantação

Autorizadores personalizados do API Gateway

Abaixo da superfície, os autorizadores personalizados Lambda para o API Gateway são usados para todas as chamadas de API (RESTful ambas WebSocket e baseadas) para validar se um determinado usuário tem permissão para realizar uma ação com base nos grupos aos quais ele pertence. Esse autorizador personalizado é apoiado por uma tabela do DynamoDB contendo as políticas de cada grupo. Ao invocar uma API, o API Gateway invoca a função Lambda do autorizador personalizado, que decodifica o token de acesso fornecido pelo Amazon Cognito para determinar a quais grupos de usuários o usuário pertence. A tabela de políticas é então consultada pelo nome do grupo para retornar a política relevante para esse grupo.

Em cada nova implantação de caso de uso, a política administrativa é atualizada para armazenar uma nova instrução que permite a ação Execute-API:invoke na API desse caso de uso. Quando os casos de uso são excluídos, a declaração correspondente é removida da política.

Para os grupos criados para um caso de uso individual, somente uma única declaração está presente na política, permitindo a ação Execute-API:invoke somente na API desse caso de uso.

Devido a essa estrutura, qualquer usuário pertencente ao grupo de um caso de uso pode acessar a API desse caso de uso. Um único usuário também pode ser adicionado manualmente a vários grupos para permitir que esse usuário use vários casos de uso.

> **Warning**
>
> Você também pode editar manualmente as políticas de um determinado grupo na tabela de políticas se quiser conceder acesso a um novo caso de uso a um grupo existente de usuários. O grupo de casos de uso é excluído quando o caso de uso é excluído (mesmo que você tenha feito edições manuais), portanto, tenha cuidado ao excluir um caso de uso.

No caso em que uma pilha de casos de uso é implantada de forma independente (sem o uso do painel de implantação), um grupo de usuários do Amazon Cognito é criado para essa implantação contendo um único usuário com acesso à API desse caso de uso. Esse grupo de usuários pertence somente a esse caso de uso e não é compartilhado entre outras implantações autônomas.

Caso de uso de texto

Suporte de streaming

Em um aplicativo de bate-papo, a latência é uma métrica importante para permitir uma experiência de usuário responsiva. A possibilidade de as inferências do LLM levarem de segundos a minutos oferece desafios sobre a melhor forma de oferecer conteúdo aos clientes. Por esse motivo, vários provedores de LLM permitem transmitir respostas de volta para o chamador. Em vez de esperar que toda a inferência seja concluída antes de retornar uma resposta, cada token pode ser retornado quando estiver disponível.

Para apoiar o uso desse recurso, o caso de uso do Text foi projetado para usar uma WebSocket API para apoiar a experiência de bate-papo. Isso WebSocket é implantado por meio do API Gateway. O uso de uma WebSocket API permite que uma conexão seja criada no início de uma sessão de bate-papo e que as respostas sejam transmitidas por esse soquete. Isso permite que os aplicativos de front-end forneçam uma melhor experiência ao usuário.

> **Note**
>
> Mesmo que um modelo forneça suporte de streaming, isso não significa necessariamente que a solução será capaz de transmitir respostas de volta por meio da WebSocket API.

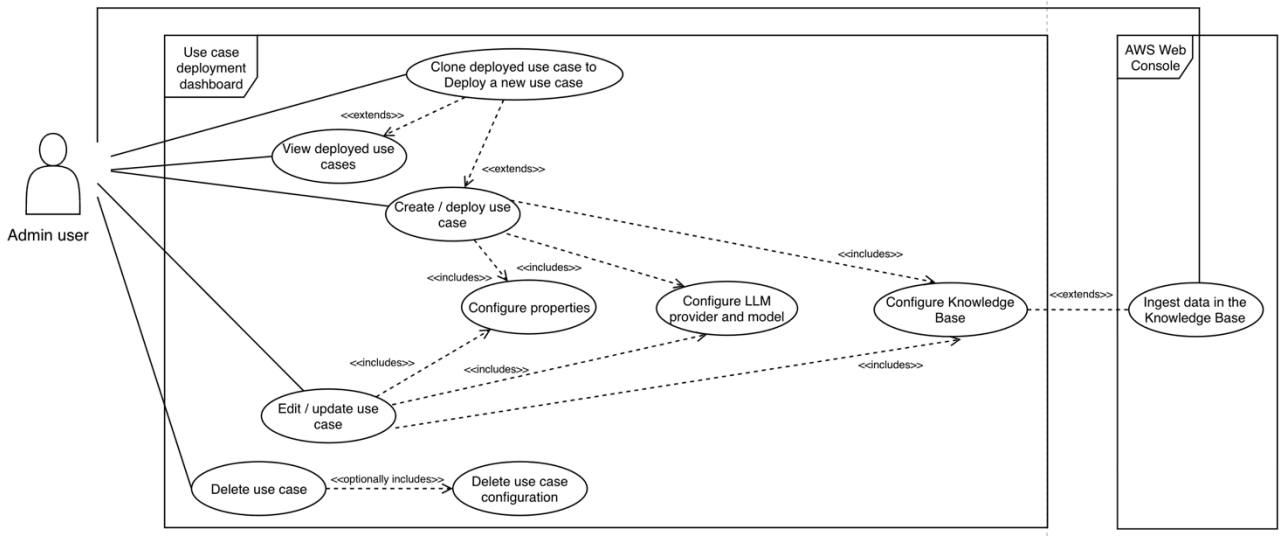
É necessário que a solução habilite a lógica personalizada para oferecer suporte ao streaming para cada provedor de modelo. Se o streaming estiver disponível, os usuários administradores poderão usar enable/disable esse recurso no momento da implantação.

Como funciona a solução Generative AI Application Builder na AWS

O usuário administrador interage principalmente com o painel de implantação para visualizar, criar e gerenciar implantações de casos de uso novos e existentes. Por meio desse painel, o usuário administrador tem acesso às seguintes ações:

- Veja a lista de implantações
- Crie novas implantações
- Editar implantações existentes
- Clone a configuração de uma implantação para criar uma nova implantação
- Excluir uma implantação (desprovisionar os recursos por meio de uma CloudFormation exclusão)
- Exclua permanentemente os detalhes de configuração de uma implantação

Descreve o diagrama de casos de uso para o usuário administrador do painel de implantação

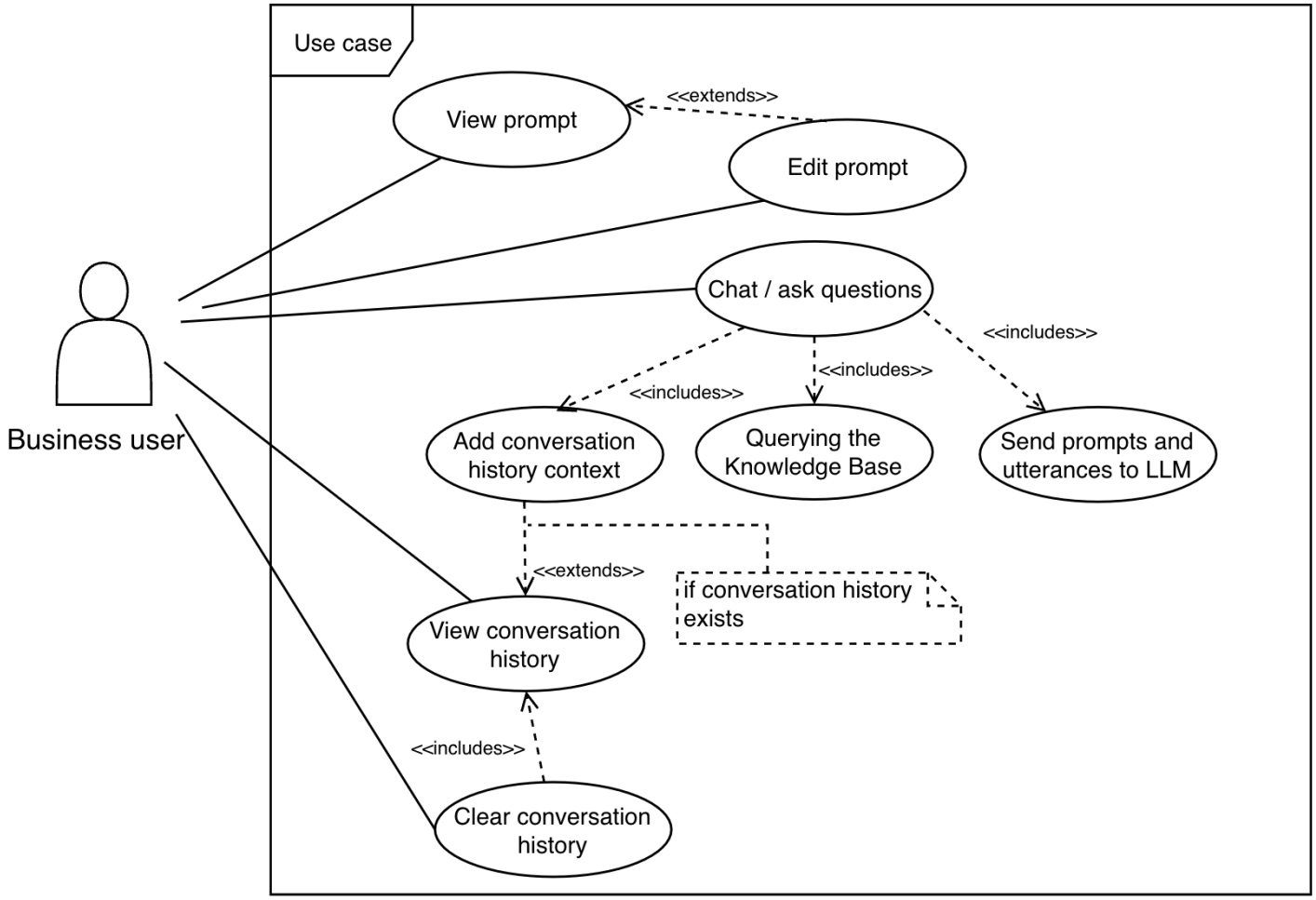


> **Note**
>
> O usuário administrador pode não ter acesso direto ao console da AWS. Nesse caso, o usuário administrador deve trabalhar com o DevOps usuário para apoiar ações como a ingestão de dados em uma base de conhecimento da Kendra.

Para o caso de uso do Text, o usuário corporativo tem acesso a uma interface de usuário que permite conversar com o LLM. As especificidades dessa configuração são controladas pelas configurações de implantação definidas pelo usuário administrador. No caso de uso do Text, o usuário corporativo tem acesso às seguintes ações:

- Envie mensagens pela interface de bate-papo
- Exibir histórico de conversas
- Limpe o histórico da conversa
- Exibir aviso
- Solicitação de edição

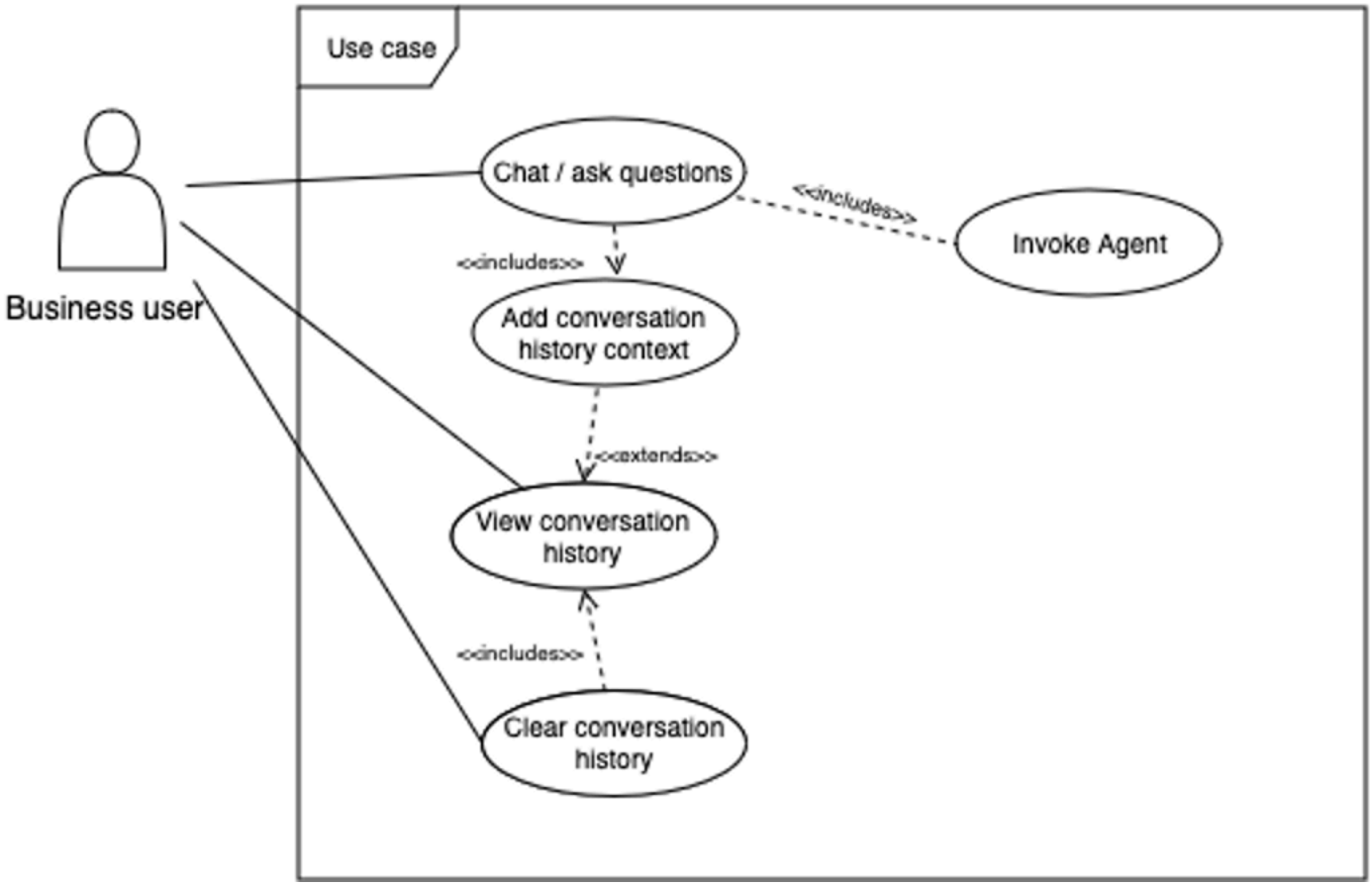
Descreve o diagrama de caso de uso para o usuário comercial do caso de uso do Text



Com o caso de uso do Bedrock Agent, o usuário corporativo pode acessar uma interface de usuário para conversar com o Amazon Bedrock Agent configurado. O usuário administrador pode definir esses detalhes nas configurações de implantação. No caso de uso do Bedrock Agent, o usuário comercial tem acesso às seguintes ações:

- Envie mensagens pela interface de bate-papo
- Exibir histórico de conversas
- Limpe o histórico da conversa

Descreve o diagrama de casos de uso para o usuário comercial do caso de uso do Bedrock Agent



Construtor de agentes

O Agent Builder fornece uma plataforma para criar, implantar e gerenciar agentes de IA prontos para produção no Amazon Bedrock. AgentCore Esta seção descreve os componentes técnicos e os detalhes da implementação.

AgentCore integração

O Agent Builder usa uma abordagem de implantação baseada em configuração com imagens de agentes pré-criadas para permitir implantações de agentes rápidas, seguras e escaláveis.

Imagens de agentes pré-criadas

As imagens do contêiner do agente são criadas pela equipe do GAAB durante o CI/CD pipeline e publicadas em um repositório ECR público. Cada versão da imagem está vinculada à versão da

solução GAAB (por exemplo, v4.0.0 →:v4.0.0). gaab-strands-agent As imagens são baseadas no SDK Strands e incluem:

- Ambiente de tempo de execução do agente
- Integração com o cliente MCP
- Capacidades de gerenciamento de memória
- OpenTelemetry instrumentação

Cache pull-Through do ECR

A solução usa o ECR Pull-Through Cache para distribuir automaticamente imagens do agente do repositório ECR público para o ECR privado do cliente. Esse serviço gerenciado pela AWS:

- Armazena imagens em cache na primeira extração (atraso de 2 a 5 minutos)
- Elimina a lógica personalizada de cópia de imagens
- Fornece disponibilidade de imagens locais para implantações subsequentes
- Cria regras de cache exclusivas por implantação para evitar conflitos

Armazenamento de configuração

As configurações do agente são armazenadas no DynamoDB junto com as configurações de casos de uso existentes. Cada configuração inclui:

- Modelo de prompt do sistema
- Fornecedor do modelo e ID do modelo
- Parâmetros do modelo (temperatura, max_tokens)
- Referências e endpoints do servidor MCP
- Configurações de memória (alternância de memória de longo prazo)
- Metadados de implantação

Registro da versão da imagem

Uma tabela do DynamoDB rastreia as versões disponíveis da imagem do agente e seu URIs cache, permitindo o gerenciamento de versões e a compatibilidade com versões anteriores.

Configuração do agente

Solicitações do sistema

Os prompts do sistema definem o comportamento, a personalidade e as capacidades do agente. Os usuários administradores podem:

- Edite o modelo padrão por meio da interface do Agent Builder
- Inclua instruções para uso da ferramenta e formatação de respostas
- Redefina para o modelo padrão a qualquer momento

Seleção de modelos

O Agent Builder oferece suporte aos modelos Amazon Bedrock na v4.0.0:

- Fornecedor de modelos: Amazon Bedrock (única opção na v4.0.0)
- Seleção de modelos: Claude, Nova e outros modelos Bedrock
- Parâmetros do modelo: temperatura, max_tokens, top_p e configurações específicas do modelo

Integração do servidor MCP

Os servidores do Model Context Protocol fornecem aos agentes acesso a ferramentas e dados corporativos:

- Descoberta de servidores por meio do endpoint da API GET /mcp
- Configuração dinâmica sem alterações de código
- Autenticação e gerenciamento de endpoints
- Capacidade da ferramenta: exposição a agentes

Streaming e processamento

Streaming em tempo real

O Agent Builder usa eventos enviados pelo servidor (SSE) de forma interligada WebSocket para streaming AgentCore de respostas em tempo real:

- A função Lambda estabelece conexão SSE com o Runtime AgentCore

- Os fluxos são vinculados ao API Gateway WebSocket
- Permite token-by-token a entrega de respostas aos clientes
- Mantém a conexão para solicitações de longa duração

Restrições de processamento

O processamento do agente na v4.0.0 está limitado ao tempo limite de execução do Lambda:

- Tempo máximo de processamento: 15 minutos
- Modelo de processamento síncrono
- Adequado para agentes conversacionais e fluxos de trabalho moderados
- Suporte assíncrono estendido planejado para a versão 4.1+

Gerenciamento de memória

Memória de curto prazo

Ativado por padrão para todos os agentes usando um personalizado MemoryHookProvider:

- Captura eventos de conversação por meio de manipuladores de retorno de chamada Strands
- Organiza por ActorID e SessionID para isolamento de contexto
- Mantém o contexto da conversa nas sessões
- Integração automática com a AgentCore memória

Memória de longo prazo

Recurso opcional usando a Ferramenta AgentCore de Memória de strands_tools:

- Simples alternância na interface do Agent Builder
- Estratégia de memória semântica com configurações padrão
- Acesso controlado por agente por meio da invocação natural de ferramentas
- Armazena insights extraídos em todas as sessões
- Usa ConversationID como SessionID

Observabilidade

OpenTelemetry Distribuição da AWS (ADOT)

Os agentes são instrumentados automaticamente durante a construção do contêiner:

- Geração automática de rastreamento para operações de agentes
- Rastreamento distribuído entre os limites do serviço
- Registro estruturado com correlação IDs
- Integração com a Pesquisa CloudWatch de Transações

Fluxo de autenticação

Os usuários se autenticam por meio do Amazon Cognito com tokens JWT validados por autorizadores personalizados do Lambda que recuperam políticas do IAM do DynamoDB com base em grupos de usuários.

Construtor de fluxo

O Workflow Builder permite a orquestração de vários agentes criando um agente supervisor que coordena vários agentes do Agent Builder usando o padrão de delegação Agents as Tools.

Arquitetura de fluxo de trabalho

Componentes-chave

- Agente supervisor: agente de ponto de entrada que recebe solicitações de usuários e delega a agentes especializados
- Agentes especializados: casos de uso do Agent Builder registrados como ferramentas para o supervisor
- Registro do agente: tabela do DynamoDB que armazena configurações e metadados do agente
- Camada de orquestração: implementação do Strands SDK do padrão Agents as Tools

Instanciação do agente

Criação de agente local

Todos os agentes especializados são instanciados localmente no mesmo AgentCore Runtime:

1. Recupera as configurações do agente do DynamoDB
2. Cria instâncias locais de cada agente do Agent Builder
3. Cada agente mantém suas próprias conexões de servidor MCP
4. Agente supervisor registra agentes especializados como ferramentas
5. O Strands SDK gerencia a seleção e delegação de agentes

Planeje a implantação

Esta seção descreve as considerações de custo, segurança, região e cota para planejar sua implantação.

> **Important**
>
> Essa solução utiliza o Amazon Bedrock como o principal serviço para acessar modelos gerados por IA. Você deve primeiro solicitar acesso aos modelos antes que eles estejam disponíveis para uso na solução. Para obter detalhes, consulte o acesso ao modelo no Guia do usuário do Amazon Bedrock.

Regiões da AWS compatíveis

> **Important**
>
> Opcionalmente, essa solução usa os serviços Amazon Bedrock e Amazon Kendra, que atualmente não estão disponíveis em todas as regiões da AWS. Você deve iniciar essa solução em uma região da AWS onde esses serviços estejam disponíveis. Para obter a disponibilidade mais atual dos serviços da AWS por região, consulte a Lista de serviços regionais da AWS.

O Generative AI Application Builder na AWS é suportado nas seguintes regiões da AWS:

Nome da região	
Leste dos EUA (Ohio)	Canadá (Central)
Leste dos EUA (Norte da Virgínia)	Europa (Frankfurt)
Oeste dos EUA (Norte da Califórnia)	Europa (Irlanda)
Oeste dos EUA (Oregon)	Europa (Londres)
Ásia-Pacífico (Mumbai)	Europa (Milão)

Nome da região	
Ásia-Pacífico (Seul)	Europa (Paris)
Ásia-Pacífico (Singapura)	Europa (Estocolmo)
Ásia-Pacífico (Sydney)	Oriente Médio (Bahrein)
Ásia-Pacífico (Tóquio)	América do Sul (São Paulo)

> **Note**
>
> Se estiver usando um modelo básico acessado fora da AWS em suas implantações, verifique com o provedor do modelo em quais regiões eles APIs estão disponíveis. Se eles APIs estiverem disponíveis apenas em determinadas regiões, você poderá experimentar instabilidade na forma de alta latência ou até mesmo tempos limite. Também é importante consultar as equipes jurídicas e de conformidade da sua organização para avaliar as considerações de que os dados ultrapassam as fronteiras regionais.

Custo

Com essa solução da AWS, você paga somente pelos recursos que usa e não há taxas mínimas nem encargos de configuração. Os usuários pagam pelo painel usado para lançar casos de uso de IA generativa e por quaisquer casos de uso implantados. O custo dos casos de uso implantados depende das configurações. Configurações de exemplo:

1. Um painel de implantação simples que custa aproximadamente $20 USD por mês.
2. Um caso de uso de chatbot simples e pronto para produção, implantado com configurações padrão em execução no Leste dos EUA (Norte da Virgínia), desenvolvido pela Amazon Bedrock sem acesso a documentos, que também custa cerca de USD 200 por mês.
3. Um sistema escalável em um caso de uso da Amazon VPC que oferece suporte a 8.000 consultas por dia em dezenas de milhares de documentos, o que custa cerca de USD 1.500 por mês. O custo do caso de uso variará dependendo da configuração, como casos de uso de texto com diferentes fornecedores de modelos, com ou sem a Geração Aumentada de Recuperação (RAG) ativada e assim por diante.

Descrição da carga de trabalho	Custo estimado (USD/mês)
Custo da amostra para o painel de implantação	$20/mês
Custos de amostra para uma prova de conceito baseada em texto (inclui painel de implantação e 1 caso de uso de texto, aproximadamente 100 interações por dia)	$40/mês
Custos de amostra para um mecanismo de consulta generativo de IA altamente escalável (Inclui painel de implantação, 1 caso de uso de texto e um índice Amazon Kendra para RAG) de até 100 mil documentos com aproximad amente 8 mil consultas por dia, com VPC habilitado	$1.500/mês
Custos de amostra para uma prova de conceito baseada em agente (Inclui painel de implantação, 1 caso de uso do Bedrock Agent com o Amazon Bedrock Knowledge Bases e o Amazon Bedrock Guardrails habilitados, aproximadamente 100 interações por dia)	$840/mês
Custos de amostra para o servidor MCP (Inclui painel de implantação, 1 caso de uso do servidor MCP com método Gateway para integração com Lambda, aproximadamente 100 invocações de ferramentas por dia)	$22/mês
Custos de amostra para o Agent Builder	$55/mês

Descrição da carga de trabalho	Custo estimado (USD/mês)
(Inclui painel de implantação, 1 caso de uso do Agent Builder com integração MCP e memória de longo prazo ativada, aproximadamente 100 interações por dia)	
Custos de amostra para o Workflow Builder (Inclui painel de implantação, 1 fluxo de trabalho com 3 agentes do Agent Builder, aproximadamente 100 interações por dia)	$109/mês

Important

Esses exemplos servem apenas para ajudá-lo a estimar os custos de suas cargas de trabalho específicas. O uso de diferentes LLMs configurações ou serviços da AWS pode alterar seus custos (por exemplo, serverless/on-demand billing vs. provisioned/time -billed). Para gerenciar custos, recomendamos criar um orçamento por meio do AWS Cost Explorer. Os preços estão sujeitos a alterações. Para obter detalhes completos, consulte a página de preços de cada serviço da AWS usado nesta solução.

Custos de amostra para executar o painel de implantação

A tabela a seguir fornece o detalhamento dos custos de um painel de implantação com parâmetros padrão e 100 usuários ativos na região Leste dos EUA (Norte da Virgínia) por um mês, o que custará cerca de USD 20/mês.

Serviço da AWS	Dimensões	Custo [USD]
API Gateway, DynamoDB, CloudFront Amazon S3, Lambda, Systems Manager Parameter Store	5.000 chamadas de API REST de 512 KB por mês sem o armazenamento em cache ativado	$1,97

Serviço da AWS	Dimensões	Custo [USD]
Amazon Cognito	100 usuários ativos por mês com recursos avançados de segurança ativados e nenhum usuário fazendo login por meio da federação SAML ou OIDC	$5,55
AWS WAF	10.000 solicitações da web em 1 ACL da web e 7 regras definidas sem nenhum grupo de regras	$12,60
Custo total do painel de implantação		$20,12

Custos de amostra para uma prova de conceito baseada em texto

Um painel de implantação pode ter muitos casos de uso implantados em um determinado momento. A tabela a seguir mostra o detalhamento dos custos de um caso de uso implantado sem RAG para 1 usuário corporativo realizando 100 consultas por dia com o LLM. As consultas são enviadas como uma mensagem de texto no WebSocket e a resposta é transmitida de volta como tokens, supondo que o streaming esteja ativado. Usando o modelo Amazon Bedrock Nova Pro, o custo de execução desse caso de uso é de cerca de $20/mês.

Serviço da AWS	Dimensões	Custo [USD]
API Gateway (WebSocket) CloudFront, Lambda, Amazon S3, AWS Systems Manager Parameter Store	100 interações de bate-papo por dia. Tamanho médio da mensagem 32 KB por mensagem e 5 minutos por conexão.	$0,61
CloudWatch	CloudWatch Registros de 1,5 GB com o modo detalhado ativado para experimentação	$7,23

Serviço da AWS	Dimensões	Custo [USD]
Amazon DynamoDB	Tabela de histórico de conversas, 1 GB de armazenamento Tabela de configuração LLM, armazenamento de 1 GB	$3,05
Subtotal dos custos do caso de uso (não incluído LLMs)		$10,89
Amazon Bedrock (Nova Pro)	Suposições para 100 interações por dia: * Custo mensal de 190 mil tokens de entrada por dia = 0,152 USD × 30 * Custo mensal de 16 mil tokens de saída por dia = 0,0512 USD × 30	$6,10
Custo total do aplicativo com o Amazon Bedrock (Nova Pro)	$10,89 (custo do caso de uso) + $6,10 (custo do Amazon Bedrock)	$17,00

Note

Os custos das chamadas de inferência feitas para serviços fora da rede da AWS não estão incluídos nessas estimativas. Consulte o guia de preços do seu provedor de LLM se você não estiver usando um provedor de modelos da AWS.
Os guias de preços dos serviços da AWS podem ser encontrados em: preços do Amazon Bedrock e preços do Amazon SageMaker AI.

Custos de amostra para um mecanismo de consulta generativo de IA altamente escalável

A tabela a seguir fornece o detalhamento dos custos de um caso de uso habilitado para RAG com o modelo Nova Pro do Amazon Bedrock como LLM. Quando uma Base de Conhecimento Bedrock é adicionada, esse caso de uso custa cerca de $1300/mês

Serviço da AWS	Dimensões	Custo [USD]
API Gateway (WebSocket)	8000 interações de bate-papo por dia. Tamanho médio da mensagem 32 KB por mensagem e 5 minutos por conexão.	$38,89
CloudFront	240.000 solicitações por mês com 100 GB de dados transferidos para a Internet e 1 GB de dados transferidos para a origem	$8,76
Amazon Bedrock (Nova Pro)	Suposições: Tokens de entrada = promptTemplate (400) + contexto (400) + chatHistory (1080) + consulta Tokens de entrada (20) = 1.900 Tokens de saída = 160 (média) Com 8.000 transações por dia, Custo diário dos tokens de entrada (1.900 x 8.000 = 15.200.000 tokens x	$487,80

Serviço da AWS	Dimensões	Custo [USD]
	0,0008/1000 de preço por token) Custo diário dos tokens de saída (160 x 8.000 = 1.280.000 tokens x 0,0032/1000 de preço por token) Custo mensal (($12,16 + $4,10) x 30)	
CloudWatch	24 métricas usando 5 GB de dados ingeridos para registros e 1 painel	$9,72
DynamoDB	Tabela do DynamoDB para acompanhar o histórico de conversas com cada registro de até 1 KB de dados, 8.000 leituras e gravações por dia	$11,70
Lambda	Tamanho do contêiner - 128 MB, 512 MB efêmero armazenamento, 2 funções Lambda usadas para autorização Tamanho do contêiner - 256 MB, 512 MB de armazenamento temporário, 5 solicitações por segundo com tempo médio de computação de 20 segundos	$20,89

Serviço da AWS	Dimensões	Custo [USD]
Custo total do caso de uso		577,76 USD/mês + custo da base de conhecimento (veja abaixo)

> **Note**
>
> Os custos das chamadas de API feitas para quaisquer serviços fora da rede da AWS não estão incluídos nessas estimativas. Consulte o guia de preços do seu provedor de LLM se não estiver usando o Amazon Bedrock.

Custos para adicionar uma base de conhecimento

Os custos da base de conhecimento variarão com base no tipo de base de conhecimento usada e (no caso da Bedrock) no armazenamento de vetores de apoio usado pela base de conhecimento. O provisionamento e o gerenciamento das bases de conhecimento estão fora do escopo da solução.

Bases de conhecimento do Amazon Bedrock

A solução não gerencia nem provisiona nenhum recurso relacionado às bases de conhecimento Amazon Bedrock. O Amazon Bedrock não incorre em custos pelo uso do recurso de base de conhecimento em si, no entanto, você será cobrado pelo uso do modelo de incorporação usado pelo seu caso de uso em cada consulta. Além disso, o armazenamento vetorial de apoio da sua base de conhecimento (por exemplo, um índice no Amazon OpenSearch Service ou um banco de dados dentro do Amazon Relational Database Service) terá um custo associado que não pode ser fornecido ou calculado aqui.

Para o cenário de mecanismo de consulta de IA generativo altamente escalável acima, os custos incorridos por esse serviço para chamar o modelo de incorporação Amazon Bedrock são os seguintes:

Serviço da AWS	Dimensões	Custo [USD]
Amazon Bedrock (Amazon Titan Text Embeddings V2)	8.000 consultas por dia com 1.900 tokens de entrada por	$9,00

Serviço da AWS	Dimensões	Custo [USD]
	consulta = 15.200.000 tokens = USD 0,30 por dia. Custo diário x 30 dias = custo mensal de $9,00 USD	
Exemplo de uso do Amazon OpenSearch Service (sem servidor)	Configuração básica sem servidor com 4 unidades de OpenSearch computação (OCU) (mínimo faturável) = USD 23,04 por dia Custo diário x 30 dias = $691,20 USD **Note** Isso fornece uma estimativa aproximada, pois algumas cargas de trabalho exigirão mais OCUs, enquanto os clientes com OpenSearch recursos provisionados existentes terão menos custos aqui.	$691,20
Custo adicional total		$700,20

Amazon Kendra

A solução pode provisionar um índice Kendra para você, ou você pode trazer o seu próprio. O custo de executar uma configuração adequada ao mecanismo de consulta generativo de IA altamente escalável acima é o seguinte:

Serviço da AWS	Dimensões	Custo [USD]
Amazon Kendra	De 0 a 8.000 consultas por dia e até 100.000 documento s com o Amazon Kendra Enterprise Edition com 0 a 50 fontes de dados	$1.008,00

> **Note**
>
> Você pode compartilhar o índice do Amazon Kendra entre os casos de uso, mas isso pode aumentar o número de consultas por índice. Se isso não estiver incluído na edição Amazon Kendra Enterprise, cobranças adicionais serão aplicadas.

Custo incremental de habilitar o Amazon VPC para um caso de uso

A tabela a seguir fornece o detalhamento dos custos de habilitar a Amazon VPC para um caso de uso implantado em dois. AZs

Serviço da AWS	Dimensões	Custo [USD]
Gateway NAT da Amazon	Suposição: implantação de 2 AZ, com um gateway NAT em cada AZ. 100 GB de dados processados por meio do NAT Gateway 730 horas, 100 GB de dados processados por mês	$74,70
AWS PrivateLink (VPC Endpoints)	Suposições: implantação de 2 AZ, com 1 sub-rede privada em cada AZ e 1 VPC Endpoint com 2 interfaces de rede elástica (). ENIs	$97,84

Serviço da AWS	Dimensões	Custo [USD]
	6 VPC endpoints, 2 por ENIs VPC endpoint, 730 horas com 1.024 GB de dados processados em um mês	
IPv4 Endereço público	Suposição: implantação de 2 AZ, 1 sub-rede pública em cada AZ com um gateway NAT em cada sub-rede pública. Cada gateway NAT configurado com 1 público IPv4 ativo. 2 IPv4 endereços públicos ativos x 730 horas em um mês x cobrança horária de 0,005 USD = 7,3 USD	$7,30
Custos adicionais (para Amazon VPC)		$179,93

Implicações de custo ao usar a taxa de transferência provisionada

Os custos de produção provisionados variarão com base no tipo de modelo que você provisionou e no período de compromisso, bem como nas unidades modelo selecionadas para o período de compromisso. Há um custo adicional associado ao uso da taxa de transferência provisionada.

Para obter mais informações e obter mais up-to-date preços, consulte os preços do Bedrock.

Custo do uso da inferência entre regiões

Não há custo adicional de roteamento ou transferência de dados para o uso da inferência entre regiões. Você paga o mesmo preço por token para modelos que em sua região de origem ou principal.

Custos de amostra para uma prova de conceito baseada em agente

Ao usar o Amazon Bedrock Agents, você é cobrado com base nos componentes que compõem o agente, como o modelo de apoio e a base de conhecimento (se o RAG estiver ativado), além de recursos adicionais que você adiciona. A tabela a seguir mostra o detalhamento de custos de um caso de uso do Bedrock Agent configurado com um modelo Claude 3.5 Sonnet sob demanda, Amazon Bedrock Knowledge Bases e Amazon Bedrock Guardrails.

Semelhante ao custo de adicionar as bases de conhecimento do Amazon Bedrock, essa solução não gerencia nem provisiona recursos relacionados aos Amazon Bedrock Agents. A solução também não gera custos para usar o Amazon Bedrock Knowledge Bases, mas tem custos para:

- Usando o modelo de incorporação para cada consulta enviada a ele
- O armazenamento vetorial de apoio para sua base de conhecimento (por exemplo, um índice no Amazon OpenSearch Service ou um banco de dados dentro do Amazon RDS)

A tabela a seguir pressupõe 100 interações por dia com 1.900 tokens de entrada e 160 tokens de saída por consulta.

> **Note**
>
> Para este exemplo de caso de uso do Bedrock Agent, se houvesse um grupo de ação configurado para usar uma API externa, esses custos seriam adicionais. Eles estão fora do escopo dos cálculos nesta tabela.

Serviço da AWS	Dimensões	Custo [USD]
API Gateway (WebSocket) CloudFront, Lambda, Amazon S3, Systems Manager Parameter Store	100 interações de bate-papo por dia, tamanho médio da mensagem 32 KB por mensagem, 5 minutos por conexão	$0,61
CloudWatch	CloudWatch Registros de 1,5 GB com o modo detalhado ativado para experimentação	$7,23

Serviço da AWS	Dimensões	Custo [USD]
DynamoDB	Tabela de configuração LLM para tamanho de registro de 1 KB e armazenamento de 1 GB	$0,25
Subtotal dos custos (não incluindo LLMs)		$8,09
Soneto antrópico Claude 3.5	* Custo diário de 190 mil tokens de entrada por dia (0,003/1.000 tokens) = $0,57 + Custo diário × 30 dias = $17,10 * Custo diário de 16 mil tokens de saída por dia (0,015/1.000 tokens) = $0,24 + Custo diário × 30 dias = $7,20	$24,30
Amazon Bedrock (Amazon Titan Text Embeddings V2) para bases de conhecimento do Amazon Bedrock	Custo diário de 190 mil tokens de entrada por dia (0,00002/ 1000 tokens) = 0,004 Custo diário × 30 dias = 0,12 USD	0,12 US$
Exemplo de uso do Amazon OpenSearch Service (sem servidor)	Configuração básica sem servidor com 4 × Unidade de OpenSearch Computação (OCU) (mínimo faturável) = $23,04 por dia Custo diário × 30 dias = $691,20	$691,20

Serviço da AWS	Dimensões	Custo [USD]
Barreiras de proteção do Amazon Bedrock	190 mil tokens equivalem aproximadamente a 760 mil (190.000 × 4) caracteres e 3.800 unidades de texto (760 mil caracteres/200) Considere uma grade de proteção configurada com filtros de conteúdo, filtro de informações de identific ação pessoal (PII), filtro de informações confidenciais (expressão regular) e filtros de palavras Custo diário do filtro de conteúdo (0,75/1000 unidades de texto) +custo do filtro de PII (0,1/1000 unidades de texto) + filtro de informações confidenciais (regex) + filtros de palavras = $2,85 + $0,38 + $0 + $0 Custo mensal = custo diário × 30 dias = $96,90	$96,90
Custo total da inscrição para um agente apoiado pelo Anthropic Claude 3.5 Sonnet	$8,09 (custo do caso de uso) + $812,52 (outras configura ções do agente)	$820,61

> **Note**
>
> Consulte o guia de preços do seu provedor de LLM se você não estiver usando um provedor de modelos da AWS. Os guias de preços dos serviços da AWS podem ser encontrados em: preços do Amazon Bedrock e preços do Amazon SageMaker AI.

Custos de amostra para o servidor MCP

Os casos de uso do MCP Server permitem a implantação e o gerenciamento de servidores do Model Context Protocol no Amazon Bedrock AgentCore. A tabela a seguir mostra o detalhamento de custos de um caso de uso do MCP Server usando o método Gateway para agrupar funções Lambda existentes.

A solução gerencia a implantação e a configuração do AgentCore Gateway. Você é cobrado por:

- Custos de infraestrutura (API Gateway, Lambda, DynamoDB, S3) CloudWatch
- AgentCore Consumo do gateway (por invocação de ferramenta)
- Custos de execução da função Lambda (para o método Gateway com alvos Lambda)
- Custos externos de API (para o método Gateway com destinos de API ou servidor MCP, se aplicável)

Item	Cálculos	Custo
Amazon API Gateway (API REST)	100 invocações de ferramentas por dia × 30 dias = 3.000 solicitações por mês	0,05 USD
AWS Lambda (orquestração)	100 invocações por dia × 30 dias × média de 1 segundo × 512 MB = 3.000 GB-segundos por mês	0,05 USD
Amazon DynamoDB	3.000 read/write solicitações por mês + 1 GB de armazenamento	0,15 US$

Item	Cálculos	Custo
Amazon CloudWatch	Monitoramento e registro padrão para 3.000 invocações	$1,00
Amazon S3	Armazenamento e registros de configuração (uso mínimo)	$0,25
Amazon Bedrock AgentCore Gateway	3.000 invocações de ferramentas por mês	0,05 USD
Função Lambda de destino	100 invocações por dia × 30 dias × 0,5 segundos × 128 MB = 1.500 GB-segundos por mês	$0,25
Custo mensal total	$1,75 (infraestrutura) + $0,05 (Gateway) AgentCore	$1,80

> **Note**
>
> Os custos variam com base no método de implantação (Gateway versus Runtime), nos tipos de destino e nos padrões de uso. As implantações do método Runtime incorrem em cobranças AgentCore de Runtime em vez de cobranças de Gateway. Os custos externos da API e os custos de hospedagem de contêineres personalizados são adicionais.

Custos de amostra para o Agent Builder

O Agent Builder permite que você crie e implante agentes personalizados no Amazon Bedrock AgentCore. A tabela a seguir mostra o detalhamento de custos de um caso de uso do Agent Builder configurado com Claude 3.5 Sonnet, integração do servidor MCP e memória de longo prazo ativada.

A solução gerencia a implantação e a configuração do AgentCore Runtime. Você é cobrado por:

- Custos de infraestrutura (API Gateway, Lambda, DynamoDB, S3) CloudWatch
- AgentCore Consumo de tempo de execução (horas de CPU e memória com base no tempo real de execução do agente)
- Inferência do modelo básico (tokens de entrada e saída)

- AgentCore Memória (eventos de curto prazo e armazenamento/recuperação de longo prazo)

A tabela a seguir pressupõe 100 interações por dia com 1.900 tokens de entrada e 160 tokens de saída por consulta, com um tempo médio de execução do agente de 5 segundos por interação.

Serviço da AWS	Dimensões	Custo [USD]
API Gateway (WebSocket) CloudFront, Lambda, Amazon S3, Systems Manager Parameter Store	100 interações de bate-papo por dia, tamanho médio da mensagem 32 KB por mensagem, 5 minutos por conexão	$0,61
CloudWatch	CloudWatch Registros de 1,5 GB com o modo detalhado ativado para experimentação	$7,23
DynamoDB	Tabela de configuração LLM para tamanho de registro de 1 KB e armazenamento de 1 GB	$0,25
Subtotal dos custos de infraestrutura		$8,09
Tempo de execução do Amazon Bedrock AgentCore	* CPU: 1 vCPU × 5 segundos × 100 interações = 125 vCPU-seconds/day = 0.140 vCPU-hours/day + Custo diário: 0,140 × 0,0895 USD = 0,013 USD + Custo mensal: 0,013 USD × 30 = 0,38 USD * Memória: 512 MB (0,5 GB) × 5 segundos × 100 interaçõe s = 250 GB- seconds/day = 0.069 GB-hours/day + Custo diário: 0,069 × 0,00945 USD =	$0,40

Serviço da AWS	Dimensões	Custo [USD]
	0,0007 USD + Custo mensal: 0,0007 USD × 30 = 0,02 USD	
Soneto antrópico Claude 3.5	* Custo diário de 190 mil tokens de entrada por dia (0,003/1.000 tokens) = 0,57 USD + Custo diário × 30 dias = 17,10 USD * Custo diário de 16 mil tokens de saída por dia (0,015/1.000 tokens) = 0,24 USD+Custo diário × 30 dias = 7,20 USD	$24,30
Memória Amazon Bedrock AgentCore	* Memória de curto prazo: 100 novos events/day × 0,25 USD/1.000 eventos = 0,025 USD/dia + Custo mensal: 0,025 USD × 30 = 0,75 USD * Armazenamento de memória de longo prazo (estratégia integrada): 100 registros × $0,75/1.000 = $0,075/mês records/month * Recuperação de memória de longo prazo: 100 retrievals/day × 0,50 USD/1.000 recuperaç ões = 0,05/dia + custo mensal: 0,05 USD × 30 = 1,50 USD	$2,33
Custo total do aplicativo para o Agent Builder com Claude 3.5 Sonnet	$8,09 (infraestrutura) + $0,40 (tempo de AgentCore execução) + $24,30 (modelo) + $2,33 (memória)	$35,12

> **Note**
>
> AgentCore O preço em tempo de execução é baseado no consumo. Os custos reais dependem de:
>
> - Tempo de execução do agente (uso de CPU e memória durante o processamento ativo)
> - Número de interações e sua complexidade
> - Uso da ferramenta MCP (adicional CPU/memory para execução da ferramenta)
> - Configuração de memória (memória de curto prazo versus memória de longo prazo ativada)
>
> Para obter AgentCore preços detalhados, consulte os preços do Amazon Bedrock.

> **Note**
>
> Se estiver usando servidores MCP que invocam serviços externos APIs ou externos, esses custos são adicionais e estão fora do escopo desse cálculo. Da mesma forma, se estiver usando ferramentas de AgentCore navegador ou intérprete de código, as cobranças baseadas no consumo se aplicam a 0,0895 USD por hora de vCPU e 0,00945 USD por GB/hora.

Custos de amostra para o Workflow Builder

O Workflow Builder cria um agente supervisor que orquestra vários agentes do Agent Builder. A tabela a seguir mostra o detalhamento dos custos de um fluxo de trabalho com 1 agente supervisor e 3 agentes especializados do Agent Builder, todos configurados com o Claude 3.5 Sonnet e com memória de longo prazo ativada.

Suposições: 100 interações por dia, média de 2 delegações de agentes por interação, 5 segundos de tempo de execução por agente.

Serviço da AWS	Dimensões	Custo [USD]
API Gateway (WebSocket) CloudFront, Lambda, Amazon	100 interações de bate-papo por dia, tamanho médio	$0,61

Serviço da AWS	Dimensões	Custo [USD]
S3, Systems Manager Parameter Store	da mensagem 32 KB por mensagem, 5 minutos por conexão	
CloudWatch	CloudWatch Registros de 1,5 GB com o modo detalhado ativado para experimentação	$7,23
DynamoDB	Tabela de configuração LLM para tamanho de registro de 1 KB e armazenamento de 1 GB	$0,25
Subtotal dos custos de infraestrutura		$8,09
Amazon Bedrock AgentCore Runtime (agente supervisor)	* CPU: 1 vCPU × 5 segundos × 100 interações = 0,140 vCPU hours/day × 30 = $0.38 * Memory: 0.5 GB × 5 seconds × 100 interactions = 0.069 GB-hours/day - × 30 = 0,02 USD	$0,40
Amazon Bedrock AgentCore Runtime (3 agentes especializados)	* Média de 2 delegações por interação = 200 agentes executions/day * CPU: 1 vCPU × 5 seconds × 200 = 0.278 vCPU-hours/day × 30 = $0.75 * Memory: 0.5 GB × 5 seconds × 200 = 0.139 GB-hours/day × 30 = 0,04 USD	$0,79
Anthropic Claude 3.5 Sonnet (agente supervisor)	* Entrada: 190K tokens/day × $0,003/1K = $0,57/dia × 30 = $17,10 * Saída: 16K × $0,015/1K = $0,24/dia × 30 = $7,20 tokens/day	$24,30

Serviço da AWS	Dimensões	Custo [USD]
Soneto antrópico Claude 3.5 (agentes especializados)	* Média de 2 delegações por interação * Entrada: 380 mil tokens/day × 0,003/mil USD = 1,14 USD/dia × 30 = 34,20 USD* Saída: 32 mil × 0,015/1 K = 0,48/dia × 30 = 14,40 USD tokens/day	$48,60
Amazon Bedrock AgentCore Memory (agente supervisor)	* Curto prazo: 100 events/da y × 0,25 USD/1.000 × 30 = 0,75 USD* Armazenamento de longo prazo: 100 registros × 0,75 USD/1K = 0,08 USD* Recuperação de longo prazo: 100 × 0,50 USD/1K × 30 = 1,50 USD retrievals/day	$2,33
Amazon Bedrock AgentCore Memory (agentes especiali zados)	* Curto prazo: 200 events/da y × 0,25 USD/1.000 × 30 = 1,50 USD* Armazenamento de longo prazo: 200 registros × 0,75 USD/1K = 0,15 USD* Recuperação de longo prazo: 200 × 0,50 USD/1K × 30 = 3,00 USD retrievals/day	$4,65
Custo total do aplicativo para o Workflow Builder com 3 agentes	$8,09 (infraestrutura) + $1,19 (tempo de AgentCore execução) + $72,90 (modelos) + $6,98 (memória)	$89,16

Note

- Taxas de delegação mais altas aumentam proporcionalmente o consumo de tokens

Para obter AgentCore preços detalhados, consulte os preços do Amazon Bedrock.

Segurança

Quando você cria sistemas na infraestrutura da AWS, as responsabilidades de segurança são compartilhadas entre você e a AWS. Esse modelo de responsabilidade compartilhada reduz sua carga operacional porque a AWS opera, gerencia e controla os componentes, incluindo o sistema operacional do host, a camada de virtualização e a segurança física das instalações nas quais os serviços operam. Para obter mais informações sobre segurança da AWS, visite Segurança da Nuvem AWS.

Usando modelos básicos no Amazon Bedrock

O Amazon Bedrock hospeda uma coleção de modelos, desde modelos Amazon Nova até outros modelos de fundação líderes (FMs). Ao usar o Amazon Bedrock, todos os modelos são hospedados na infraestrutura da AWS. Isso significa que, ao usar o Amazon Bedrock como provedor de LLM, todas as suas solicitações de inferência permanecerão na rede da AWS e o tráfego da rede não sairá da sua região.

Note

Todos os modelos de fundação (FMs) disponíveis por meio do Amazon Bedrock são hospedados diretamente na infraestrutura da AWS gerenciada e de propriedade da AWS. Os fornecedores de modelos não têm acesso aos dados do cliente, como solicitações e continuações, ou registros de serviços do Amazon Bedrock. Para obter informações adicionais sobre a postura de segurança do Amazon Bedrock, consulte Proteção de dados no Amazon Bedrock no Guia do usuário do Amazon Bedrock.

Perfis do IAM

As funções do IAM permitem que os clientes atribuam políticas e permissões de acesso granulares a serviços e usuários na nuvem da AWS. Essa solução cria funções do IAM que concedem às funções Lambda da solução acesso para criar recursos regionais.

CloudWatch Registros

Você pode ativar o modo detalhado ao implantar um caso de uso usando a página de seleção do modelo Deployment Dashboard, em Configurações adicionais. O modo detalhado permite CloudWatch registros detalhados que podem ser úteis para depuração e experimentação imediata.

> **Note**
>
> Quando o modo detalhado estiver ativado, os documentos recuperados da base de conhecimento (se o RAG estiver ativado) e os prompts também serão registrados, os quais podem conter informações confidenciais.

VPC

A solução oferece duas opções para a configuração da Amazon VPC:

1. Deixe a solução criar uma Amazon VPC para você.
2. Gerenciando e trazendo sua própria Amazon VPC para uso na solução.

Deixe a solução criar uma Amazon VPC para você

Se você selecionar a opção de permitir que a solução crie uma Amazon VPC, ela será implantada como uma arquitetura 2-AZ por padrão com um intervalo CIDR 10.10.0.0/20. Você tem a opção de usar o Amazon VPC IP Address Manager (IPAM), com 1 sub-rede pública e 1 sub-rede privada em cada AZ. A solução cria gateways NAT em cada uma das sub-redes públicas e configura as funções Lambda para criá-las nas sub-redes privadas. ENIs Além disso, essa configuração cria tabelas de rotas e suas entradas, grupos de segurança e suas regras, rede ACLs, endpoints VPC (endpoints de gateway e interface).

Gerenciando sua própria Amazon VPC

Ao implantar a solução com uma Amazon VPC, você tem a opção de usar uma Amazon VPC existente em sua conta e região da AWS. Recomendamos que você disponibilize sua VPC em pelo menos duas zonas de disponibilidade para garantir a alta disponibilidade. Sua VPC também deve ter os seguintes endpoints de VPC e suas políticas de IAM associadas para suas configurações de VPC e tabela de rotas.

Para um painel de implantação Amazon VPC

1. Endpoint de gateway para DynamoDB.
2. Endpoint de gateway para S3.
3. Endpoint de interface para CloudWatch.
4. Endpoint de interface para AWS CloudFormation.

Para um caso de uso: Amazon VPC

1. Endpoint de gateway para DynamoDB.
2. Endpoint de gateway para S3.
3. Endpoint de interface para CloudWatch.
4. Ponto final de interface para Systems Manager Parameter Store.

> **Note**
>
> A solução requer apenas`com.amazonaws.region.ssm`.

5. Endpoint de interface para Amazon Bedrock (bedrock-runtime, agent-runtime,). bedrock-agent-runtime
6. Opcional: se a implantação usar o Amazon Kendra como base de conhecimento, será necessário um endpoint de interface para o Amazon Kendra.
7. Opcional: se a implantação usar qualquer LLM no Amazon Bedrock, será necessário um endpoint de interface para o Amazon Bedrock.

> **Note**
>
> A solução requer apenas`com.amazonaws.region.bedrock-runtime`.

8. Opcional: se a implantação usar o Amazon SageMaker AI para o LLM, será necessário um endpoint de interface para o Amazon SageMaker AI.

> **Note**
>
> A solução não excluirá nem modificará a configuração da VPC ao usar a opção de implantação Bring your own VPC. No entanto, ele excluirá tudo o VPCs que for criado pela

solução na opção Criar uma VPC para mim. Por esse motivo, você deve ter cuidado ao compartilhar uma VPC gerenciada pela solução entre pilhas/implantações.
Por exemplo, a implantação A usa a opção Criar uma VPC para mim. A implantação B usa Traga minha própria VPC usando a VPC criada pela implantação A. Se a implantação A for excluída antes da implantação B, a implantação B não funcionará mais porque a VPC foi excluída. Além disso, como a implantação B está usando as funções ENIs criadas pelas Lambda, a exclusão da implantação A pode causar erros e retenção de recursos residuais.

Amazon CloudFront

Essa solução implanta um console web hospedado em um bucket do Amazon S3. Para ajudar a reduzir a latência e melhorar a segurança, essa solução inclui uma CloudFront distribuição com uma identidade de acesso de origem, que é um CloudFront usuário que fornece acesso público ao conteúdo do bucket do site da solução. Para obter mais informações, consulte Restringir o acesso ao conteúdo do Amazon S3 usando uma identidade de acesso de origem no CloudFront Amazon Developer Guide.

> **Note**
>
> CloudFront tem um limite de cota flexível em nível de conta de 20 políticas de cabeçalho de resposta. Essa solução cria políticas de cabeçalho de resposta personalizadas para fins de segurança. Se você tiver mais de 20 implantações do Generative AI Application Builder na AWS ou em seus casos de uso, novas implantações podem falhar devido ao limite da cota.

Para resolver esse problema, você pode solicitar um aumento de cota para a cota de políticas de cabeçalho de resposta no console de Quotas de Serviços da AWS seguindo estas etapas:

1. Abra o console de Quotas de Serviços da AWS.
2. No painel de navegação, escolha AWS services (produtos da AWS).
3. Pesquise e selecione Amazon CloudFront.
4. Role até a cota Políticas do cabeçalho de resposta e escolha Solicitar aumento da cota.
5. Siga as instruções para solicitar um aumento no limite de cota da sua conta da AWS.

Ao aumentar a cota de políticas de cabeçalho de resposta, você pode garantir que novas implantações do Generative AI Application Builder na AWS ou em seus casos de uso não falhem devido ao limite da cota.

Cotas

Service quotas, ou limites, representam o máximo de recursos ou operações de serviço permitidos em uma conta AWS.

Cotas para serviços da AWS nesta solução

Verifique se você tem cota suficiente para cada um dos serviços implementados nessa solução. Para obter mais informações, consulte as cotas de serviços da AWS.

Use os links a seguir para acessar a página desse serviço. Para visualizar as cotas de serviço para todos os serviços da AWS na documentação sem alternar páginas, consulte as informações na página Endpoints e cotas de serviços no PDF.

Cotas do Amazon Bedrock AgentCore

Para implantações do Agent Builder, esteja ciente das seguintes cotas de AgentCore serviço Amazon Bedrock:

Quota	Leste dos EUA (Norte da Virgínia)	Outras regiões
Cargas de trabalho de sessão ativa por conta	1000	500
Total de agentes por conta	1.000	1.000
Versões por conta	1.000	1.000

Implante a solução

Essa solução usa CloudFormation modelos e pilhas da AWS para automatizar sua implantação. O CloudFormation modelo especifica os recursos da AWS incluídos nessa solução e suas propriedades. A CloudFormation pilha provisiona os recursos descritos no modelo.

Visão geral do processo de implantação

Antes de lançar a solução, analise o custo, a arquitetura, a segurança e outras considerações discutidas neste guia.

> **Important**
>
> Se você planeja usar o Amazon Bedrock, você deve solicitar acesso aos modelos antes que eles estejam disponíveis para uso. Consulte o acesso ao modelo no Guia do usuário do Amazon Bedrock para obter mais detalhes.

Tempo de implantação: Aproximadamente 10 minutos

Etapa 1: iniciar a pilha do painel de implantação

Etapa 2: implantar um caso de uso

Etapa 3: implantar um caso de uso usando o assistente do painel de implantação

Etapa 4: configuração pós-implantação

Opcionalmente, você pode implantar os casos de uso separadamente da solução, se preferir não ter a interface do usuário do painel de implantação ou APIs.

- Implantação de um caso de uso de texto independente
- Implantação de um caso de uso autônomo do Bedrock Agent

Você também pode fornecer uma configuração de chat do DynamoDB.

> **Important**
>
> Essa solução envia métricas operacionais para a AWS (os “Dados”) sobre o uso dessa solução. Usamos esses dados para entender melhor como os clientes usam essa solução e os serviços e produtos relacionados. A coleta desses dados pela AWS está sujeita à Política de Privacidade da AWS.

CloudFormation Modelo da AWS

Você pode baixar o CloudFormation modelo dessa solução antes de implantá-la.

View template

generative-ai-application-builder-on-aws.template - Use esse modelo para iniciar a solução e todos os componentes associados. A configuração padrão implanta as soluções principais e de suporte encontradas nos serviços da AWS nesta seção de soluções, mas você pode personalizar o modelo para atender às suas necessidades específicas.

> **Note**
>
> Os CloudFormation recursos da AWS são criados a partir de construções do AWS Cloud Development Kit (AWS CDK).

Este CloudFormation modelo da AWS implanta o Generative AI Application Builder na AWS na nuvem da AWS.

Etapa 1: iniciar a pilha do painel de implantação

Siga as step-by-step instruções nesta seção para configurar e implantar a solução em sua conta.

Tempo de implantação: Aproximadamente 10 minutos

1. Faça login no AWS Management Console e selecione o botão para iniciar o `generative-ai-application-builder-on-aws.template` CloudFormation modelo.

   Launch solution

2. Por padrão, esse modelo é iniciado na região Leste dos EUA (Norte da Virgínia). Para iniciar a solução em outra região da AWS, use o seletor de Região na barra de navegação do console.

> **Note**
>
> Essa solução usa o Amazon Kendra e o Amazon Bedrock, que atualmente não estão disponíveis em todas as regiões da AWS. Se estiver usando esses recursos, você deve iniciar essa solução em uma região da AWS onde esses serviços estejam disponíveis. Para obter a disponibilidade mais atual por região, consulte a Lista de serviços regionais da AWS.

3. Na página Criar pilha, verifique se o URL de modelo correto é apresentado na caixa de texto URL do Amazon S3 e escolha Avançar.
4. Na página Especificar detalhes da pilha, atribua um nome para a sua pilha de soluções. Para obter informações sobre limitações de nomes de caracteres, consulte Limites do IAM e do STS no Guia do usuário do AWS Identity and Access Management.
5. Em Parâmetros, revise os parâmetros do modelo dessa solução e modifique-os conforme requerido. Esta solução usa os seguintes valores padrão.

Parâmetro	Padrão	Description
E-mail do usuário administrador	No	O endereço de e-mail do usuário administrador que terá acesso ao painel de implantação. Se fornecidos, um grupo e um usuário do Amazon Cognito serão criados com permissões para implantar e gerenciar casos de uso. Você também pode usar `placeholder@example.com` para criar o Grupo, mas não o Usuário. Consulte Configuração manual do grupo de usuários para obter informaçõ

Parâmetro	Padrão	Description
		es sobre como configurar seu grupo de usuários.
VpcEnabled	No	O painel de implantação deve ser implantado em uma VPC?
CreateNewVpc	No	Disponível somente, se VpcEnabledestiverYes. Se o valor forYes, a pilha criará a VPC e implantará a solução dentro da VPC criada. Se VpcEnabledfor Yes e CreateNewVpcforNo, você deverá fornecer uma configuração de VPC existente (ExistingVpcId,, ExistingPrivateSubnetIdsExistingSecurityGroupIds, VpcAzs).
IPAMPoolId	(Entrada opcional)	Você pode configurar o IPAM e fornecer o ID criado como entrada para atribuir o intervalo de endereços IP que a implantação dessa pilha deve usar. Para obter detalhes sobre o IPAM, consulte Como funciona o IPAM.

Parâmetro	Padrão	Description
Implantar UI	`Yes`	Você tem a opção de implantar o painel de implantação sem a interface de usuário da web (e os recursos da AWS necessári os para a implantação na web). Nesse caso, a solução implantará toda a infraestr utura, incluindo endpoints da API REST. Essa opção é útil para integrar sua própria interface da web ao painel de implantação APIs.
ExistingVpcId	(Entrada opcional)	Exigido somente se você quiser implantar a solução em uma VPC existente que você criou.
ExistingPrivateSubnetIds	(Entrada opcional)	Exigido somente se você quiser implantar a solução em uma VPC existente que você criou. As funções Lambda serão implantadas nessa sub-rede.
ExistingSecurityGroupIds	(Entrada opcional)	Exigido somente se você quiser implantar a solução em uma VPC existente que você criou. Certifique-se de que os grupos de segurança tenham as permissões para uma conexão TCP de saída.

Parâmetro	Padrão	Description
VpcAzs	(Entrada opcional)	Exigido somente se você quiser implantar a solução em uma VPC existente que você criou.
CognitoDomainPrefix	(Entrada opcional)	Obrigatório somente se você quiser implantar a solução em um grupo de usuários existente do Amazon Cognito que você criou. Se você não fornecer um valor, a solução o gerará.
ExistingCognitoUserPoolId	(Entrada opcional)	Obrigatório somente se você quiser implantar a solução em um grupo de usuários existente do Amazon Cognito que você criou.
ExistingCognitoUserPoolClient	(Entrada opcional)	Obrigatório somente se você quiser implantar a solução em um grupo de usuários existente do Amazon Cognito que você criou. Se você não fornecer um valor, a solução cria um cliente de grupo de usuários. Esse parâmetro só pode ser fornecido se você fornecer um ExistingCognitoUserPoolIdvalor.

6. Escolha Avançar.
7. Na página Configurar opções de pilha, selecione Avançar.
8. Na página Revisar e criar, revise e confirme as configurações. Selecione a caixa confirmando que o modelo criará recursos do AWS Identity and Access Management (IAM).

9. Escolha Enviar para implantar a pilha.

Você pode ver o status da pilha no CloudFormation console da AWS na coluna Status. Você deve receber o status CREATE_COMPLETE em aproximadamente 10 minutos.

Etapa 2: implantar um caso de uso

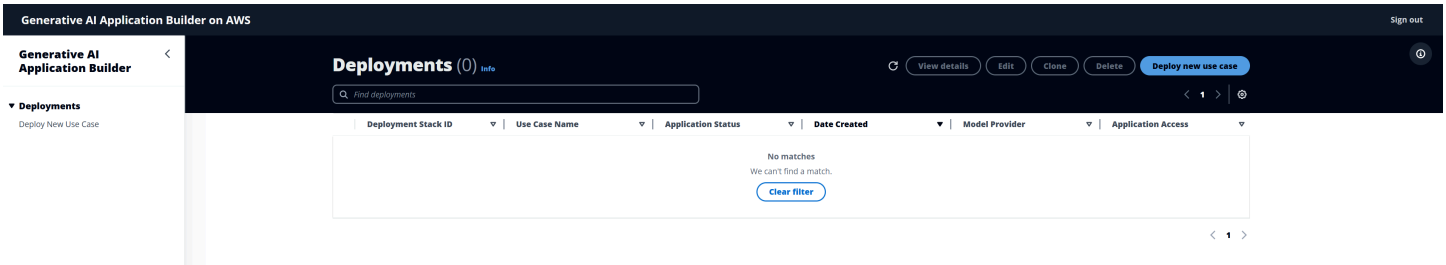
> **Important**
>
> Depois que a pilha for implantada com sucesso, um e-mail de inscrição será enviado para o e-mail do usuário administrador configurado. Usando essas credenciais, o usuário administrador pode entrar no painel de implantação para usar o aplicativo web.

> **Note**
>
> O DevOps usuário com acesso ao AWS Management Console deve fornecer ao usuário administrador a CloudFront URL da interface do usuário do painel de implantação quando a pilha for concluída. O URL pode ser encontrado na guia Saídas da CloudFormation pilha.

1. Faça login no painel de implantação como usuário administrador.
2. Na página inicial do aplicativo, escolha Implantar novo caso de uso.

   Isso inicia o assistente de implantação, que orienta você na criação do caso de uso.

Descreve a página inicial do painel de implantação - nova implantação



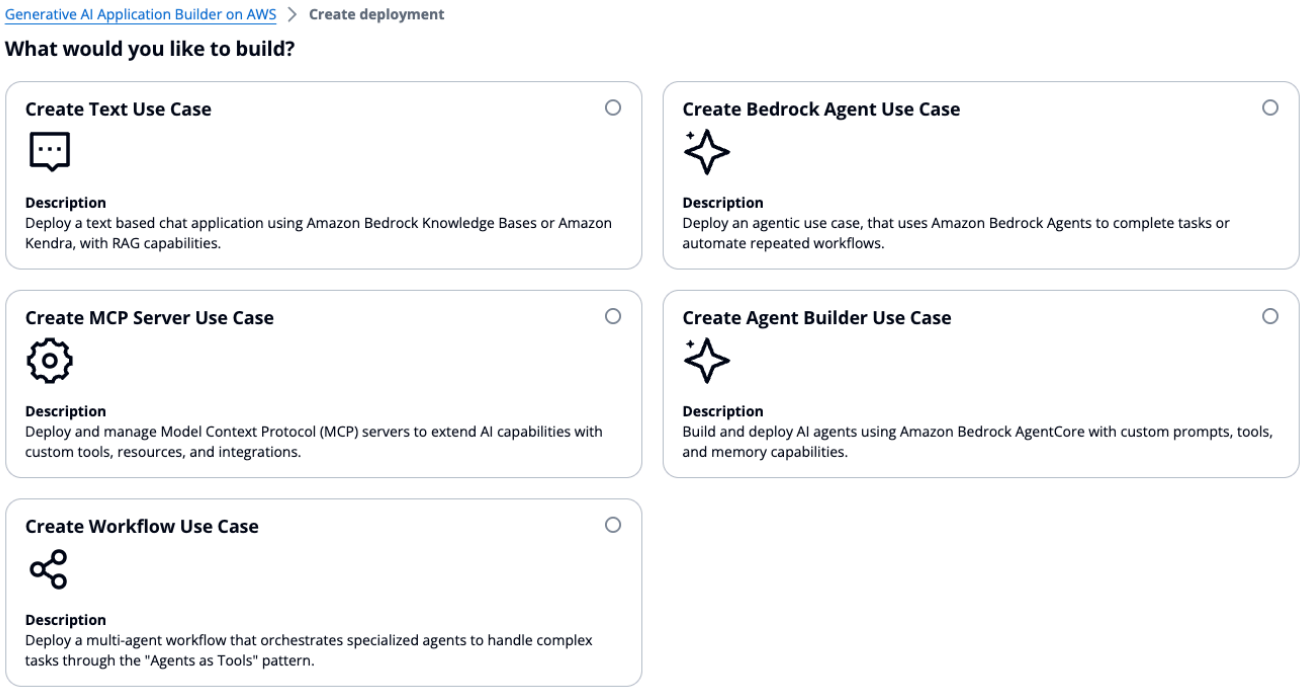
> **Note**
>
> Se você precisar adicionar mais usuários à sua implantação, consulte o grupo de usuários do Gerenciando o Cognito para obter mais detalhes.

Etapa 3: implantar um caso de uso usando o assistente do painel de implantação

No assistente do painel de implantação, você deve escolher entre as seguintes opções:

- Caso de uso de texto - implanta um aplicativo de bate-papo, com recursos de RAG opcionais
- Caso de uso do Bedrock Agent - usa Amazon Bedrock Agents para concluir tarefas ou automatizar fluxos de trabalho repetidos
- Servidor MCP - Implemente e gerencie servidores MCP com métodos de gateway ou tempo de execução
- Agent Builder - Crie e implante agentes personalizados AgentCore com integração MCP e gerenciamento de memória
- Construtor de fluxo de trabalho - orquestre vários agentes do Agent Builder usando delegação hierárquica

Mostra cinco opções: Criar caso de uso do Text, Criar caso de uso do Bedrock Agent, Criar caso de uso do MCP Server, Criar caso de uso do Agent Builder ou Criar caso de uso do fluxo de trabalho.



Etapa 3a: implantar um caso de uso de texto

Esta seção fornece instruções para implantar um caso de uso do Text.

Selecione o caso de uso

Quando você escolhe Criar caso de uso de texto, a interface do usuário abre a tela Selecionar caso de uso. Forneça as informações a seguir:

- Nome do caso de uso.
- Endereço de e-mail opcional para que o usuário padrão do caso de uso seja adicionado ao grupo de usuários do Amazon Cognito para o caso de uso e receba permissões para interagir com ele.
- Se você deseja implantar uma interface de usuário com esse caso de uso. Se você não quiser implantar uma interface de usuário com o caso de uso, você pode usar os endpoints de API implantados para uso com seu aplicativo.

Detalhes do caso de uso

A etapa de detalhes do caso de uso permite que você defina configurações adicionais para sua implantação.

Por padrão, o caso de uso do Text cria e configura um grupo de usuários do Amazon Cognito para você quando a solução implanta o painel de implantação. A solução autentica novos casos de uso com um cliente recém-criado no mesmo grupo de usuários. No entanto, você pode fornecer uma ID de grupo de usuários e uma ID de cliente existentes nesta etapa se quiser usar seu próprio grupo de usuários e cliente do Amazon Cognito com o caso de uso.

Important

Os usuários administradores têm acesso a todos os casos de uso implantados quando o grupo de usuários do Amazon Cognito é criado por meio do assistente de implantação. Se você fornecer seu próprio grupo de usuários durante a implantação, deverá garantir que o administrador tenha as permissões para acessar os casos de uso implantados.
Você também precisará atualizar o retorno de chamada permitido URLs e o desligamento permitido URLs em seus clientes de aplicativos no Cognito. Para fazer isso:

1. Navegue até o console do Cognito
2. Escolha Grupos de usuários.
3. Escolha seu grupo de usuários.
4. Escolha Clientes de aplicativos no menu à esquerda.
5. Escolha o cliente do aplicativo que você deseja modificar.
6. Escolha a guia Páginas de login.
7. Escolha Editar e adicione seu URLs.
8. Escolha Salvar alterações.

Além disso, se você precisar adicionar mais usuários a um caso de uso, consulte a seção Gerenciando o grupo de usuários do Cognito.

Selecione a configuração de rede

Essa etapa do assistente permite que você implante o caso de uso com uma Amazon Virtual Private Cloud (Amazon VPC) pré-existente ou nova. Se selecionar uma VPC preexistente, você deverá fornecer uma ID de VPC, até 16 IDs de sub-rede e até 5 grupos de segurança IDs para usar com essa VPC. Se você não estiver usando uma VPC preexistente, essas configurações serão definidas para você.

Criar o modelo

Na etapa Selecionar modelo, você pode escolher o fornecedor do modelo no menu suspenso. Existem duas opções: Bedrock e. SageMaker

Se você selecionar SageMaker, poderá criar um endpoint do modelo de SageMaker IA no console de SageMaker IA e fornecer o esquema de entrada que o modelo espera e a saída JSONPath para a resposta do LLM. Você pode consultar a seção Usando a Amazon SageMaker AI como um provedor de LLM e os exemplos de carga útil de SageMaker IA fornecidos no repositório da GitHub solução.

Se você selecionar Amazon Bedrock, você verá quatro opções:

- Modelos de início rápido - Comece rapidamente com uma coleção de modelos com price/performance características diferentes. Recomendado para criar seus primeiros aplicativos. Essa opção permite que você selecione um nome de modelo na lista fornecida.
- Outros modelos de fundação - Acesse a gama completa de modelos de fundação com diferentes capacidades e especializações. Essa opção permite que você insira o ID do modelo de base sob demanda do Bedrock desejado.
- Perfis de inferência — Os perfis de inferência utilizam a inferência entre regiões da Bedrock para aumentar a taxa de transferência e melhorar a resiliência, roteando suas solicitações em várias regiões da AWS durante picos de utilização. Essa opção permite que você insira a ID do perfil de inferência que você deseja usar.
- Modelos provisionados - Capacidade de transferência dedicada para cargas de trabalho de produção que exigem desempenho consistente. Essa opção permite que você insira o ARN do provisioned/custom modelo a ser usado no Amazon Bedrock.

A etapa de seleção do modelo também permite que você escolha as configurações avançadas do modelo. Consulte as configurações avançadas do LLM para obter detalhes sobre a configuração do Amazon Bedrock Guardrails, a taxa de transferência provisionada para o Amazon Bedrock e parâmetros adicionais do modelo.

Inferência entre regiões

A inferência entre regiões ajuda os usuários do Amazon Bedrock a gerenciar facilmente picos de tráfego não planejados usando a computação em diferentes regiões da AWS. Para usar a inferência entre regiões, você precisa do perfil de inferência. Um perfil de inferência é uma abstração sobre um pool de recursos sob demanda de um conjunto configurado de regiões da AWS. Ele pode rotear sua solicitação de inferência, originada da sua região de origem, para outra região configurada nesse

pool. Isso permite a distribuição do tráfego em várias regiões da AWS. Isso ajuda a permitir maior produtividade e maior resiliência durante períodos de pico de demanda.

Os perfis de inferência são nomeados de acordo com o modelo e as regiões que eles suportam. Você deve chamar um perfil de inferência de uma das regiões que ele inclui. Por exemplo, conforme mostrado na tabela a seguir, o ID do perfil de inferência `us.anthropic.claude-3-haiku-20240307-v1:0` permite a distribuição do tráfego `us-east-1` e das `us-west-2` regiões do modelo escolhido. Alguns modelos só estão disponíveis com um perfil de inferência em uma região específica.

Perfil de inferência	ID do perfil de inferência	Regiões incluídas
US Anthropic Claude 3 Haiku	`us.anthropic.claude-3-haiku-20240307-v1:0`	Leste dos EUA (Norte da Virgínia) (`us-east-1`) Oeste dos EUA (Oregon) (`us-west-2`)

Se você quiser usar uma ID de perfil de inferência em vez de uma ID de modelo, deverá identificar a ID de perfil de inferência apropriada. Consulte Regiões e modelos compatíveis para perfis de inferência no Guia do usuário do Amazon Bedrock para obter mais informações. No console do Amazon Bedrock, a opção de inferência entre regiões no menu de navegação à esquerda fornece esses perfis de inferência. IDs

Depois de identificar o ID do perfil de inferência a ser usado, você pode usá-lo durante o estágio Selecionar modelo executando as seguintes etapas:

1. Selecione Amazon Bedrock como fornecedor do modelo.
2. Selecione a opção de botão de opção Perfis de inferência.
3. Insira o ID do seu perfil de inferência na caixa de texto que aparece.

Consulte Melhorar a resiliência com inferência entre regiões no Guia do usuário do Amazon Bedrock para obter mais detalhes sobre perfis de inferência.

Selecione a base de conhecimento

Se você deseja implantar um caso de uso de geração aumentada sem recuperação (RAG), pode pular esta etapa.

No entanto, se você deseja habilitar o RAG como parte de sua implantação, agora você pode fornecer um Amazon Kendra Index Id pré-configurado ou um ID da Base de Conhecimento Amazon Bedrock. Você também pode criar um novo Amazon Kendra Index para uso com a solução. Atualmente, a solução oferece suporte às bases de conhecimento Amazon Kendra e Amazon Bedrock como bases de conhecimento para sua implantação de casos de uso baseados em RAG.

Consulte a seção Configurando uma base de conhecimento para obter diretrizes sobre a ingestão de dados na base de conhecimento para uso com sua implantação baseada em RAG.

Configurações avançadas de RAG

O assistente permite que você selecione opções avançadas para uso com a implantação do RAG, como o número de documentos a serem recuperados sempre que uma consulta é enviada à sua base de conhecimento, uma resposta de texto estático do LLM quando nenhum documento é encontrado na base de conhecimento, se você deseja exibir fontes de documentos com sua resposta do LLM para verificações de integridade, etc. Além disso, você também pode configurar configurações específicas da base de conhecimento para o Amazon Kendra, como Controle de Acesso Baseado em Funções (RBAC) ou Substituir Tipo de Pesquisa ao usar o Amazon Serverless com o Amazon Bedrock Knowledge Bases. OpenSearch Consulte a seção Configurações avançadas da Base de Conhecimento para obter mais detalhes sobre essas configurações avançadas.

> **Note**
>
> Sua base de conhecimento deve estar na mesma conta e região do painel de implantação implantado e das pilhas de casos de uso.

Selecione solicitações e limites de token

Nesta etapa, você pode configurar seu prompt para uso com o LLM. Os prompts podem exigir espaços reservados`{input}`, como e. `{history}` `{context}` Esses espaços reservados instruem o LLM sobre onde extrair as informações do usuário, o histórico de conversas e as informações recuperadas da base de conhecimento.

- Para o fornecedor do modelo Bedrock, o prompt do sistema deve ser fornecido, sem restrições para um caso de uso que não seja do RAG. O aviso de desambiguação para o fornecedor do modelo Bedrock, no entanto, requer um mínimo de dois espaços reservados - e `{input}` `{history}`

- Para solicitações de fornecedor de SageMaker modelos, sistema e desambiguação, ambas exigem no mínimo dois espaços reservados - e. `{input}` `{history}`
- Para casos de uso do RAG, para cada fornecedor de modelo, o `{context}` espaço reservado também é necessário.

Para obter mais informações, consulte Configurando seus prompts. Você também pode consultar a seção Dicas para gerenciar os limites de tokens do modelo ao selecionar os tamanhos dos limites de tokens para suas solicitações.

Ativar entrada multimodal

Essa etapa permite que você habilite recursos de entrada multimodais para seu caso de uso. Quando ativado, os usuários podem carregar e enviar imagens e documentos junto com suas consultas de texto.

Tipos de arquivos e restrições compatíveis:

- Imagens: até 20 imagens por mensagem. Cada imagem não deve ter mais de 3,75 MB de tamanho e 8.000 px de altura e largura. Formatos suportados: png, jpeg, gif, webp
- Documentos: até 5 documentos por mensagem. Cada documento não deve ter mais do que 4,5 MB de tamanho. Formatos suportados: pdf, csv, doc, docx, xls, xlsx, html, txt, md

Como usar a entrada multimodal:

1. Ative o MultimodalEnabledparâmetro durante a implantação do caso de uso
2. Na interface de bate-papo, os usuários podem fazer upload de arquivos de duas maneiras:
   - Clicar no botão de upload na caixa de entrada do bate-papo ou
   - Arrastar e soltar arquivos diretamente na interface de bate-papo
3. Os arquivos são enviados para o Amazon S3 e processados pelo modelo selecionado
4. Os arquivos enviados são excluídos automaticamente após 48 horas

Rastreamento do status do arquivo:

DevOps os usuários podem monitorar os metadados do arquivo no DynamoDB, o que inclui o tempo de upload e o status do processamento. Os arquivos podem ter os seguintes status:

- pendente - o upload do arquivo foi iniciado, mas ainda não foi concluído. Esse é o status inicial quando uma URL pré-assinada é gerada.
- carregado - O arquivo foi carregado com sucesso para o S3 e está pronto para ser processado pelo modelo.
- excluído - O arquivo foi excluído pelo usuário e não deve mais estar acessível para processamento.
- inválido - Verificações de validação de arquivo com falha (por exemplo, incompatibilidade de tipo de arquivo ou falha na validação de segurança).

Arquivos com status pendente que nunca foram enviados serão limpos automaticamente quando o TTL expirar. Somente arquivos com status de upload podem ser processados pelo modelo.

O bucket multimodal do S3 e a tabela de metadados do DynamoDB estão disponíveis nas saídas do Deployment Dashboard com as chaves e, respectivamente. `MultimodalDataBucketName` `MultimodalDataMetadataTable`

> **Note**
>
> Nem todos os modelos suportam entrada multimodal. Verifique se o modelo selecionado oferece suporte ao processamento de imagens e documentos antes de ativar esse recurso. Consulte os modelos de base suportados na documentação do Amazon Bedrock para verificar quais modelos oferecem suporte à imagem como modalidade de entrada.

> **Important**
>
> Os arquivos enviados pelos usuários são armazenados no Amazon S3 com uma política de ciclo de vida de 48 horas. Os metadados sobre os arquivos enviados são armazenados no Amazon DynamoDB com um TTL de 24 horas para o histórico de conversas.

Revise e implante

Após essa etapa, revise as configurações selecionadas e escolha Implantar caso de uso. O novo caso de uso então é implantado e fica visível na visualização do painel de implantação para gerenciar ainda mais.

Etapa 3b: Implantar um caso de uso do Bedrock Agent

O caso de uso do Bedrock Agent fornece um mecanismo poderoso e seguro para invocar os Amazon Bedrock Agents em seus casos de uso. Esse recurso permite que os desenvolvedores integrem perfeitamente os recursos de agentes autônomos baseados em IA que podem orquestrar e executar tarefas de várias etapas em vários modelos básicos, fontes de dados, aplicativos de software e conversas com usuários, mantendo medidas de segurança robustas.

Pré-requisitos

Antes de criar um agente Amazon Bedrock, verifique se você tem o seguinte:

1. A conta da AWS na qual o Generative AI Application Builder na AWS é implantado, com acesso ao console Amazon Bedrock.
2. Permissões apropriadas do IAM para criar e gerenciar Amazon Bedrock Agents.

Criação de um agente Amazon Bedrock

Consulte Criar e configurar o agente manualmente no Guia do usuário do Amazon Bedrock para obter instruções detalhadas sobre como criar um agente. Você pode configurar opções como:

- Instruções (solicitações) para seu agente
- Base de conhecimento, usada para pesquisar informações adicionais com base na entrada do usuário
- Memória do agente para permitir que os agentes lembrem informações em várias sessões (por no máximo 30 dias)

Depois de criar com sucesso um agente do Amazon Bedrock, você pode prosseguir para o fluxo do assistente de casos de uso do Generative AI Application Builder no AWS Bedrock Agent. Para fazer isso, escolha Implantar um novo caso de uso no painel de implantação e selecione Criar caso de uso do Bedrock Agent. Siga o assistente e use as etapas a seguir para configurar o caso de uso.

Selecione o caso de uso

Essa etapa é a mesma do caso de uso de texto descrito anteriormente.

Selecione a configuração de rede

Essa etapa é a mesma do caso de uso de texto descrito anteriormente

Selecione o agente

Nesta etapa, você deve fornecer o ID do agente e o ID do alias do agente Amazon Bedrock que você criou.

Etapa 3c: Implantar um caso de uso do MCP Server

O caso de uso do servidor MCP (Model Context Protocol) permite que você implante e gerencie servidores MCP que podem ser integrados a modelos e agentes de IA. Os servidores MCP fornecem uma forma padronizada de expor ferramentas, recursos e capacidades aos aplicativos de IA. Você pode criar servidores MCP a partir de funções APIs Lambda existentes ou hospedar servidores MCP personalizados usando imagens de contêiner.

Pré-requisitos

Antes de implantar um caso de uso do MCP Server, verifique se você tem o seguinte:

1. A conta da AWS na qual o Generative AI Application Builder na AWS é implantado.
2. Permissões apropriadas do IAM para criar e gerenciar AgentCore recursos do Amazon Bedrock.
3. Dependendo do método de criação escolhido:
   - Para o método Gateway (Lambda/API/MCPservidor): funções Lambda, endpoints de API com seus arquivos de esquema correspondentes (formato JSON para Lambda OpenAPI/Smithy , APIs for) ou endpoints de URL do servidor MCP
   - Para o método Runtime (ECR): uma imagem de contêiner Docker enviada para o Amazon ECR contendo sua implementação de servidor MCP

Métodos de criação do MCP Server

A solução oferece suporte a dois métodos para criar servidores MCP:

Crie a partir do Lambda, API ou MCP Server (método Gateway)

Esse método cria um gateway MCP que envolve funções Lambda existentes, REST ou servidores MCP externos APIs, tornando-os acessíveis como ferramentas MCP. O gateway lida com a tradução de protocolos entre o MCP e seus serviços existentes.

- Destinos do Lambda: integre as funções existentes do Lambda fornecendo o ARN da função e um arquivo de esquema JSON descrevendo o formato da função input/output

- Destinos da OpenAPI: integre o REST usando as especificações APIs da OpenAPI (formato JSON ou YAML) com suporte para 2.0 ou autenticação de chave de API OAuth
- Alvos do Smithy: integre os APIs definidos usando arquivos de modelo do Smithy (formato.smithy ou.json)
- Objetivos do MCP Server: conecte-se diretamente a servidores MCP externos por meio de endpoints de URL, permitindo a integração de servidores MCP existentes sem implantar uma nova infraestrutura

Você pode configurar vários destinos (até 10) em um único gateway MCP, cada um representando uma ferramenta ou recurso diferente.

Hospedagem a partir do ECR Image (método Runtime)

Esse método implanta um servidor MCP em contêiner a partir de uma imagem do Amazon ECR. Use essa abordagem quando você tiver uma implementação de servidor MCP personalizada que precisa ser executada como um serviço independente.

- Forneça o URI da imagem ECR (deve incluir uma tag, por exemplo, `:latest` ou `:v1.0.0`)
- Opcionalmente, configure variáveis de ambiente para passar a configuração para seu contêiner
- O contêiner deve implementar o protocolo MCP e expor os endpoints necessários

Implantando um servidor MCP

Para implantar um caso de uso do MCP Server, escolha Implantar um novo caso de uso no painel Deployment e selecione Create MCP Server Use Case. Siga o assistente e use as etapas a seguir para configurar o caso de uso.

Selecione o caso de uso

Essa etapa é a mesma do caso de uso de texto descrito anteriormente.

Selecione a configuração de rede

Atualmente, somente o acesso público está habilitado e o VPC não é compatível com a configuração da rede.

Criar servidor MCP

Nesta etapa, você configura a implantação do servidor MCP:

Método de criação do servidor MCP

Escolha entre os dois métodos de criação:

- Crie a partir do Lambda, API ou MCP Server: crie um gateway MCP a partir de funções Lambda existentes, especificações de API ou endpoints de servidor MCP externos
- Hospedagem a partir da imagem ECR: implante um servidor MCP personalizado a partir de uma imagem de contêiner

> **Note**
>
> O método de criação não pode ser alterado após a implantação. Se você precisar alternar métodos, deverá implantar um novo caso de uso do MCP Server.

Configuração do gateway (para Lambda/API/MCP o método Server)

Se você selecionou o método Gateway, configure um ou mais destinos:

1. Nome do alvo (obrigatório): um nome amigável para identificar essa configuração de destino
2. Descrição do alvo (opcional): uma breve descrição do que esse alvo faz
3. Tipo de alvo: selecione o tipo de alvo a ser configurado:
   - Lambda: Para funções do AWS Lambda
   - OpenAPI: Para REST com especificações APIs OpenAPI
   - Smithy: Para APIs com definições do modelo Smithy
   - Servidor MCP: Para conexão direta com servidores MCP externos por meio de endpoints de URL
4. Arquivo do esquema (obrigatório): faça o upload do arquivo do esquema que descreve seu destino:
   - Para Lambda: arquivo de esquema JSON descrevendo o formato. input/output Para obter detalhes sobre a criação de esquemas de ferramentas Lambda, consulte Esquema da ferramenta Lambda no Amazon Bedrock Developer Guide. AgentCore
   - Para a OpenAPI: arquivo de especificação da OpenAPI (JSON ou YAML). Para obter detalhes sobre os requisitos do esquema OpenAPI, consulte o esquema OpenAPI no Amazon Bedrock Developer Guide. AgentCore

- Para Smithy: arquivo de modelo Smithy (.smithy ou .json). Para obter detalhes sobre a criação de metas do Smithy, consulte Criação de metas do Smithy no Guia do desenvolvedor do Amazon Bedrock. AgentCore

5. ARN da função Lambda (necessário para destinos Lambda): o ARN da função Lambda a ser integrada
6. URL do servidor MCP (necessário para destinos do servidor MCP): O ponto final do URL do servidor MCP externo ao qual se conectar. O URL deve ser codificado corretamente e o servidor MCP deve oferecer suporte aos recursos da ferramenta com as versões do protocolo MCP 2025-06-18. Para obter mais informações, consulte os destinos dos servidores MCP no Amazon Bedrock AgentCore Developer Guide.
7. Autenticação de saída (necessária para destinos OpenAPI): configure a autenticação para chamadas de API REST:
   - Tipo de autenticação: escolha OAuth 2.0 ou chave de API
   - ARN do provedor de autenticação de saída: o ARN do provedor de credenciais no cofre de tokens Amazon Bedrock AgentCore
   - Configurações adicionais: dependendo do tipo de autenticação:
     - Para OAuth 2.0: configure escopos e parâmetros personalizados
     - Para chave de API: especifique a localização (cabeçalho ou parâmetro de consulta), nome do parâmetro e prefixo opcional

Você pode adicionar vários alvos (até 10) escolhendo Adicionar outro alvo. Cada destino representa uma ferramenta ou recurso separado exposto pelo seu servidor MCP.

Configuração ECR (para o método de imagem ECR)

Se você selecionou o método Runtime, forneça:

1. URI da imagem do ECR (obrigatório): o URI completo da sua imagem do Docker no Amazon ECR
   - Formato: `account-id.dkr.ecr.region.amazonaws.com/repository-name:tag`.
   - A imagem deve estar na mesma região da AWS da sua implantação
   - É necessária uma tag (por exemplo,`:latest`,`:v1.0.0`)
2. Variáveis de ambiente (opcional): configure pares de valores-chave para passar para seu contêiner em tempo de execução
   - Use-os para fornecer configuração, credenciais ou sinalizadores personalizados

- Você pode adicionar até 10 variáveis de ambiente

Revise e implante

Depois de configurar seu servidor MCP, revise as configurações selecionadas e escolha Deploy Use Case. O novo caso de uso do MCP Server então é implantado e fica visível na visualização do painel de implantação para gerenciamento adicional.

> **Note**
>
> As implantações do MCP Server criam recursos no Amazon Bedrock AgentCore, incluindo gateways, tempos de execução e identidades de carga de trabalho. Esses recursos são gerenciados automaticamente pela solução e serão limpos quando você excluir o caso de uso.

Etapa 3: implantar um caso de uso do Agent Builder

O Agent Builder permite que você crie, configure e implante agentes de IA prontos para produção no Amazon Bedrock. AgentCore Esse recurso fornece controle total sobre o comportamento do agente por meio de solicitações do sistema, seleção de modelos, integração do servidor MCP e gerenciamento de memória.

O processo de implantação é basicamente o mesmo de um caso de uso de texto, com algumas diferenças notáveis.

Selecione o caso de uso

Essa etapa é a mesma do caso de uso de texto descrito anteriormente.

Detalhes do caso de uso

Essa etapa é a mesma do caso de uso de texto descrito anteriormente.

Configurar agente

Nesta etapa, você define as configurações do agente principal, incluindo o prompt do sistema, as servers/Strands ferramentas MCP disponíveis e a memória.

Prompt do sistema

O prompt do sistema define o comportamento, a personalidade e as capacidades do agente. Você pode:

- Edite o modelo padrão de prompt do sistema
- Use o botão Redefinir para o padrão para restaurar o modelo original
- Inclua instruções para uso da ferramenta e formatação de respostas

Integração do servidor MCP (opcional)

Configure os servidores do Model Context Protocol para fornecer ao seu agente acesso às ferramentas e dados corporativos:

1. Selecione entre os servidores MCP disponíveis no menu suspenso
2. Analise as ferramentas prontas para uso que estarão acessíveis ao agente

> **Note**
>
> Os servidores MCP devem estar configurados e acessíveis antes da implantação. Consulte a documentação do MCP para obter instruções de configuração do servidor.

Configuração de memória

Configure como o agente mantém o contexto e o conhecimento:

- Memória de curto prazo: ativada por padrão para todos os agentes. Mantém o contexto da conversa nas sessões.
- Memória de longo prazo: alterne para permitir a extração e o armazenamento de insights em todas as sessões. Usa AgentCore memória com estratégia de memória semântica.

Revise e implante

Após essa etapa, revise as configurações selecionadas e escolha Implantar caso de uso. A implantação do Agent Builder normalmente é concluída em 10 a 15 minutos. O novo caso de uso então se torna visível na visualização do painel de implantação para gerenciar ainda mais.

Etapa 3e: Implantar um caso de uso de fluxo de trabalho

O Construtor de fluxo de trabalho permite que você crie agentes supervisores que orquestram vários agentes do Agent Builder usando o padrão de delegação de Agentes como Ferramentas. Esse recurso permite que você crie fluxos de trabalho multiagentes complexos reutilizando as implantações existentes do Agent Builder.

O processo de implantação segue um padrão semelhante ao Agent Builder, com etapas adicionais para descoberta e seleção de agentes.

Selecione o caso de uso

Essa etapa é a mesma do caso de uso de texto descrito anteriormente.

Detalhes do caso de uso

Essa etapa é a mesma do caso de uso de texto descrito anteriormente.

Configurar o agente supervisor

Nesta etapa, você configura o agente supervisor que coordenará os agentes especializados do Agent Builder.

Prompt do sistema

O prompt do sistema define como o agente supervisor delega o trabalho a agentes especializados. Você pode:

- Edite o modelo padrão de prompt do sistema
- Inclua instruções para seleção e delegação de agentes
- Defina como agregar resultados de vários agentes
- Use o botão Redefinir para o padrão para restaurar o modelo original

> **Note**
>
> O prompt do sistema deve descrever claramente quando e como usar cada agente especializado. As descrições dos agentes são essenciais para uma delegação adequada.

Seleção de modelos

Selecione o modelo básico para o agente supervisor. O agente supervisor usa esse modelo para:

- Entenda as solicitações dos usuários
- Selecione agentes especializados apropriados
- Coordene a execução do agente
- Agregar e formatar respostas

Selecione agentes especializados

Nesta etapa, você seleciona a quais agentes do Agent Builder o supervisor pode delegar trabalho.

Adicionando agentes

1. Clique em Adicionar agente para abrir a caixa de diálogo de seleção do agente
2. Selecione um ou mais agentes do Agent Builder na lista
3. Revise as descrições do agente que serão fornecidas ao supervisor
4. Confirme a seleção

> **Note**
>
> - Os fluxos de trabalho exigem pelo menos 1 caso de uso do Agent Builder como agente especializado
> - Todos os agentes especializados devem ser implantados com sucesso antes de criar o fluxo de trabalho

Revise e implante

Analise a configuração do fluxo de trabalho, incluindo:

- Solicitação e modelo do sistema do agente supervisor
- Lista de agentes especializados
- Memory Settings

Escolha Implantar caso de uso. A implantação do fluxo de trabalho normalmente é concluída em 15 a 20 minutos. O novo fluxo de trabalho se torna visível na visualização do painel de implantação para gerenciar ainda mais.

Etapa 4: configuração pós-implantação

Esta seção fornece recomendações para configurar a solução após a implantação.

Controle de versão do bucket Amazon S3, políticas de ciclo de vida e replicação entre regiões

Essa solução não impõe configurações de ciclo de vida nos buckets que ela cria. Recomendamos o seguinte:

- Definir configurações de ciclo de vida para implantações de produção. Para obter detalhes, consulte Definir a configuração do ciclo de vida em um bucket no Guia do usuário do Amazon Simple Storage Service.
- Habilitando o controle de versão e a replicação entre regiões para buckets do Amazon S3 com base no caso de uso para o qual a solução foi implantada.

Backups do Amazon DynamoDB

Essa solução usa o DynamoDB para várias finalidades (consulte os serviços da AWS nesta solução). A solução não permite backups das tabelas que ela cria. Recomendamos criar um backup desse recurso para implantações de produção. Consulte Backup de uma tabela do DynamoDB e Uso do AWS Backup for DynamoDB para obter detalhes.

CloudWatch Painel e alarmes da Amazon

A solução implanta um painel personalizado CloudWatch para renderizar gráficos de métricas personalizadas publicadas e métricas de serviços da AWS. Recomendamos criar CloudWatch alarmes e adicionar notificações com base no caso de uso para o qual a solução foi implantada.

CloudWatch Registros da Amazon

Os registros do Lambda são configurados para nunca expirar e os registros do API Gateway são configurados com uma expiração de 10 anos. Você pode atualizar a expiração dos respectivos grupos de registros para se alinhar à política de retenção de registros da sua empresa.

Domínios da web personalizados com certificados TLS v1.2 ou superior

A solução implanta uma interface de usuário da web e um Edge Optimized API Gateway usando CloudFront. CloudFrontO domínio de não impõe certificados TLS v1.2 ou superior. Recomendamos criar um domínio personalizado usando o Amazon Route 53, criar um certificado usando o AWS Certificate Manager ou usar um certificado existente, se sua organização tiver um.

Para obter detalhes adicionais, consulte o Guia do desenvolvedor do Amazon Route 53 e Escolha de uma versão mínima do TLS para um domínio personalizado no API Gateway.

Escalabilidade com o Amazon Kendra

Essa solução fornece a capacidade de usar o Amazon Kendra para realizar pesquisas inteligentes baseadas em NLP nos documentos ingeridos. Você pode aumentar a capacidade do Amazon Kendra usando os CloudFormation seguintes parâmetros para cargas de trabalho maiores:

Parâmetro	Padrão	Description
Capacidade adicional de consulta do Amazon Kendra	`0`	A quantidade de capacidad e extra de consulta para um índice e GetQuerySuggestion scapacidade. Uma unidade de capacidade adicional para um índice fornece aproximad amente 8.000 consultas por dia.
Capacidade de armazenam ento adicional do Amazon Kendra	`0`	A quantidade de capacidade de armazenamento extra para um índice. Uma unidade de capacidade única fornece 30 GB de espaço de armazenam ento ou 100.000 documentos, o que ocorrer primeiro.
Edição Amazon Kendra	`Developer`	O Amazon Kendra fornece as edições Developer e Enterpris e para criar índices. Para

Parâmetro	Padrão	Description
		obter mais informações sobre as diferenças entre as edições Amazon Kendra, consulte os preços do Amazon Kendra.

Para modificar os valores desses CloudFormation parâmetros, selecione os valores apropriados no momento da implantação da pilha. Para obter mais informações sobre unidades de consulta e capacidade de armazenamento, consulte Ajustando a capacidade.

> **Note**
>
> Se o caso de uso do Text não for implantado com o RAG habilitado, um índice do Amazon Kendra não será usado ou criado.

Configurando o SSO usando a federação Idp

Essa solução permite a integração com provedores de identidade externos que oferecem suporte à federação de identidade baseada em SAML ou OIDC. Quando a solução é implantada, ela cria um pool de usuários do Amazon Cognito e uma integração individual de clientes de aplicativos para o painel de implantação e casos de uso individuais. Com base no Idp externo, siga as etapas fornecidas na seção Configuração de provedores de identidade para seu grupo de usuários do Guia do Desenvolvedor do Amazon Cognito e escolha a integração do cliente do aplicativo para o painel de implantação ou o caso de uso com o qual você gostaria de configurar o SSO.

Para passar as informações do grupo de usuários para a base de conhecimento ou armazenamentos vetoriais em uma arquitetura baseada em RAG, você precisará mapear grupos de usuários do Idp externo para grupos de usuários do Amazon Cognito. A solução fornece um gatilho inicial da função Lambda de andaime a ser mapeado com a fase de pré-geração do token. A função Lambda tem o arquivo group_mapping.json que deve ser atualizado para fornecer os mapeamentos do grupo. Consulte Personalização dos fluxos de trabalho do grupo de usuários com gatilhos Lambda para ver os acionadores Lambda compatíveis com o Amazon Cognito.

Configuração manual do grupo de usuários

Se você optar por não enviar um e-mail de administrador ou usuário padrão durante a implantação, deverá criar manualmente os grupos de usuários apropriados no Amazon Cognito para garantir as permissões corretas:

1. Para o painel de implantação, crie um grupo chamado `Admin` em seu grupo de usuários do Cognito.
2. Para cada caso de uso, crie um grupo chamado `${UseCaseName}-Users` em seu grupo de usuários do Cognito, onde `${UseCaseName}` está o nome do seu caso de uso implantado.

Esses grupos são necessários para que o mecanismo de autorização funcione corretamente. Todos os usuários aos quais você deseja conceder acesso devem ser adicionados aos grupos apropriados.

Se `placeholder@example.com` for aprovado, o grupo Cognito será criado, mas você ainda deverá criar os usuários associados e atribuí-los ao grupo.

Personalizando a tela de login

Essa solução usa a interface de usuário hospedada pelo Amazon Cognito para renderizar a página de login. Para personalizar a página de login integrada, consulte Personalização das páginas integradas de login e cadastro no Guia do Desenvolvedor do Amazon Cognito.

Considerações adicionais sobre segurança

Com base no caso de uso para o qual você implanta a solução, revise as seguintes recomendações de segurança:

- Chaves de criptografia do AWS KMS gerenciadas pelo cliente — A solução usa chaves do AWS KMS gerenciadas pela AWS por padrão, pois elas estão disponíveis sem custo adicional. Analise seu caso de uso para determinar se você deve atualizar a solução para usar chaves do AWS KMS gerenciadas pelo cliente.
- Regras de limitação do API Gateway - A solução é implantada com regras de limitação padrão no API Gateway. Com base no seu caso de uso e nos volumes de transações esperados, recomendamos que você configure a limitação para o. APIs Para obter detalhes, consulte Solicitações da API Throttle para obter uma melhor taxa de transferência no Guia do desenvolvedor do Amazon API Gateway.

- Habilitando a AWS CloudTrail — Como prática de segurança recomendada, considere habilitar a AWS CloudTrail na conta da AWS em que a solução está implantada para registrar chamadas de API na conta da AWS. Para obter detalhes, consulte o Guia CloudTrail do usuário da AWS.
- Detecção de desvios - recomendamos configurar a detecção de desvios em CloudFormation pilhas para identificar e ser notificado sobre alterações não intencionais ou maliciosas na pilha de soluções implantadas. Para obter detalhes, consulte Implementação de um alarme para detectar automaticamente o desvio nas CloudFormation pilhas da AWS.
- Cognito JSON Web Tokens (JWTs) - A solução usa tokens emitidos pelo Amazon Cognito JWTs para se autenticar com os endpoints da API REST. Configuramos a solução com uma expiração de cinco minutos para tokens de ID e tokens de acesso. Quando um usuário se desconecta, sua capacidade de gerar novos tokens é revogada (o token de atualização é revogado). No entanto, até a expiração do token atual, todas as solicitações para o endpoint da API serão autenticadas com sucesso, pois elas têm um token válido. Analise as considerações de segurança para seu caso de uso e ajuste o período de validade do token.

Personalização de políticas de ciclo de vida:

Para implantações de produção, revise e ajuste as políticas de ciclo de vida com base em seus requisitos de retenção. Consulte Definir a configuração do ciclo de vida em um bucket no Guia do usuário do Amazon Simple Storage Service.

Armazenamento de arquivos e ciclo de vida multimodais

Se você habilitou recursos de entrada multimodais (MultimodalEnableddefinidos comoYes) para seu caso de uso, a solução cria um bucket do Amazon S3 para armazenar arquivos enviados e uma tabela do DynamoDB para rastrear os metadados do arquivo.

Políticas de ciclo de vida padrão:

- Arquivos S3: excluídos automaticamente após 48 horas
- Metadados do DynamoDB: os registros expiram após 24 horas (TTL do histórico de conversas)

Considerações de segurança:

- Os arquivos são particionados por ID de caso de uso, ID de usuário, ID de conversa e ID de mensagem e, em vez disso, um arquivo é armazenado com um nome UUID. O mapeamento do UUID para nomes de arquivos está disponível na tabela de metadados do DynamoDB

- Os usuários só podem acessar os arquivos que eles enviaram em suas próprias conversas
- A validação do tipo de arquivo é realizada usando a detecção mágica de números
- Recomendamos ativar o Amazon GuardDuty Malware Protection for S3 para verificar se há conteúdo malicioso nos arquivos enviados.

Implantando um caso de uso de texto independente

Siga as step-by-step instruções nesta seção para configurar e implantar a solução em sua conta.

Tempo de implantação: aproximadamente 10 a 30 minutos

1. Faça login no AWS Management Console e selecione o botão para iniciar o CloudFront modelo que você deseja implantar.

BedrockChat.modelo	
SageMakerChat.modelo	

2. Por padrão, esse modelo é iniciado na região Leste dos EUA (Norte da Virgínia). Para iniciar a solução em outra região da AWS, use o seletor de Região na barra de navegação do console.

   Observação: essa solução usa o Amazon Kendra e o Amazon Bedrock, que atualmente não estão disponíveis em todas as regiões da AWS. Se estiver usando esses recursos, você deve iniciar essa solução em uma região da AWS onde esses serviços estejam disponíveis. Para obter a disponibilidade mais atual por região, consulte a Lista de serviços regionais da AWS.

3. Na página Criar pilha *, verifique se o URL do modelo correto está na caixa de texto *URL do Amazon S3 *e escolha *Avançar.
4. Na página *Especificar detalhes da pilha*, atribua um nome à sua pilha de soluções. Para obter informações sobre limitações de nomes de caracteres, consulte Limites do IAM e do STS no Guia do usuário do AWS Identity and Access Management.
5. Em Parâmetros, revise os parâmetros do modelo dessa solução e modifique-os conforme requerido. Esta solução usa os seguintes valores padrão.

UseCaseUUID	*<_Requires input_>*	36 caracteres UUIDv4 para identificar esse caso de uso implantado em um aplicativo.
UseCaseConfigRecordKey	*<_Requires input_>*	Chave correspondente ao registro contendo as configurações exigidas pelo provedor de bate-papo Lambda em tempo de execução. O registro na tabela deve ter um atributo-chave correspon dente a esse valor e um atributo de configuração contendo a configuração desejada. Esse registro será preenchido pela plataform a de implantação se estiver em uso. Para implantaç ões autônomas desse caso de uso, é necessária uma entrada criada manualmen te na tabela UseCaseCo nfigTableNamedefinida em.
UseCaseConfigTableName	*<_Requires input_>*	A pilha lerá a configuração da tabela com esse nome na chave UseCaseConfigRecor dKey

ExistingRestApiId	(Entrada opcional)	ID da API REST da API Gateway existente a ser usada. Se não for fornecida , uma nova API REST do API Gateway será criada. Normalmente fornecido durante a implantação a partir do painel de implantação. Observação: usar o APIs Existing pode ajudar a reduzir a duplicação de recursos e simplificar o gerenciamento de APIs quando você precisa implantar vários casos de uso autônomos. Ao fornecer o existente APIs para um caso de uso independente, você é responsável por garantir que a API seja configurada com as rotas necessárias com os modelos esperados. Uma rota /details pré-configurada necessária (busca detalhes do caso de uso durante o bate-papo) e, opcionalm ente, uma rota /feedback (se FeedbackEnabledestiver definida como para permitir a coleta de feedback `Yes` para respostas de bate-papo do LLM) deve ser configura da. Além disso ExistingA piRootResourceId,, ExistingC ognitoUserPoolIde também

		ExistingCognitoGroupPolicyTableNamedeve ser fornecido.
ExistingApiRootResourceId	(Entrada opcional)	ID de recurso raiz da API Gateway REST existente a ser usado. O ID do recurso raiz da API REST pode ser obtido no console da AWS selecionando o recurso raiz (/) na seção “Recursos ” da API. A ID do recurso será então exibida no painel de detalhes do recurso. Como alternativa, você pode executar uma chamada de descrição da API em sua API REST para encontrar o ID do recurso raiz.
FeedbackEnabled	No	Se definido como Não, a pilha de casos de uso implantada não terá acesso ao recurso de feedback.
ExistingModelInfoTableName	(Entrada opcional)	Nome da tabela do DynamoDB para a tabela que contém informaçõ es e padrões do modelo. Usado pela plataforma de implantação. Se omitida, uma nova tabela será criada para abrigar os padrões do modelo.

DefaultUserEmail	`placeholder@example.com`	E-mail do usuário padrão para esse caso de uso. Um usuário do Amazon Cognito para esse e-mail é criado para acessar o caso de uso. Se não for fornecido, o Grupo e o Usuário do Cognito não serão criados. Você também pode usar `placeholder@example.com` para criar o Grupo, mas não o Usuário. Consulte Configuração manual do grupo de usuários para obter informações sobre como configurar seu grupo de usuários.
ExistingCognitoUserPoolId	(Entrada opcional)	UserPoolIdde um grupo de usuários existente do Amazon Cognito com o qual esse caso de uso será autenticado. Normalmente é fornecido durante a implantação a partir do painel de implantação, mas pode ser omitido ao implantar essa pilha de casos de uso de forma independente.
CognitoDomainPrefix	(Entrada opcional)	Insira um valor se quiser fornecer um domínio para o Cognito User Pool Client. Se você não fornecer um valor, a implantação gerará um.

ExistingCognitoUserPoolClient	(Entrada opcional)	Forneça um cliente de grupo de usuários (App Client) para usar um existente. Se você não fornecer um cliente de grupo de usuários, um novo será criado. Esse parâmetro só pode ser fornecido se um ID de grupo de usuários existente for fornecido.
ExistingCognitoGroupPolicyTableName	(Entrada opcional)	Nome da tabela do DynamoDB contendo políticas de grupos de usuários. Isso é usado pelo autorizador personalizado na API do caso de uso. Normalmente, você pode fornecer uma entrada ao implantar a partir da plataforma de implantação, mas pode ser omitida ao implantar essa pilha de casos de uso de forma independente.
RAGEnabled	`true`	Se definida como verdadeira, a pilha de casos de uso implantada usa o índice Amazon Kendra fornecido, criado para fornecer a funcionalidade RAG. Se definido como`false`, o usuário interage diretamente com o LLM.

KnowledgeBaseType	Bedrock	Tipo de base de conhecimento a ser usado para RAG. Defina apenas se RAGEnabled estiver`true`. Pode ser Bedrock ou Kendra. Nota: Só é relevante se RAGEnabledfor verdade.
ExistingKendraIndexId	(Entrada opcional)	ID do índice de um índice Kendra existente a ser usado para o caso de uso. Se nenhum for fornecido e KnowledgeBaseTypefor Kendra, um novo índice será criado para você. Nota: Só é relevante se RAGEnabledé `true` e Knowledge BaseTypeé`Kendra`.
NewKendraIndexName	(Entrada opcional)	Nome do novo índice Kendra a ser criado para esse caso de uso. Só se aplica se não ExistingKendraIndexIdfor fornecido. Nota: Só é relevante se RAGEnabledfor verdade e KnowledgeBaseTypefor Kendra.

NewKendraQueryCapacityUnits	0	Unidades adicionais de capacidade de consulta para o novo índice Amazon Kendra a serem criadas para esse caso de uso. Só se aplica se não ExistingKendraIndexIdfor fornecido, consulte CapacityUnitsConfiguration. Nota: Só é relevante se RAGEnabledé `true` e Knowledge BaseTypeé`Kendra`.
NewKendraStorageCapacityUnits	`0`	Unidades adicionais de capacidade de armazenamento para o novo índice Amazon Kendra a serem criadas para esse caso de uso. Só se aplica se não ExistingKendraIndexIdfor fornecido, consulte CapacityUnitsConfiguration. Nota: Só é relevante se RAGEnabledé `true` e Knowledge BaseTypeé`Kendra`.

NewKendraIndexEdition	(Entrada opcional)	A edição do Amazon Kendra a ser usada para o novo índice Amazon Kendra a ser criado para esse caso de uso. Só se aplica se não ExistingKendraInde xIdfor fornecido, consulte as edições Amazon Kendra. Nota: Só é relevante se RAGEnabledé `true` e Knowledge BaseTypeé`Kendra`.
BedrockKnowledgeBaseId	(Entrada opcional)	ID da base de conhecime nto fundamental a ser usada em um caso de uso do RAG. Não podem ser fornecidos se ExistingKendraIndexIdou NewKendraIndexNameforem fornecidos. Nota: Só é relevante se RAGEnabledé `true` e Knowledge BaseTypeé`Bedrock`.
VpcEnabled	`No`	Os recursos das pilhas devem ser implantados em uma VPC.
CreateNewVpc	`No`	Selecione`Yes`, se quiser que a solução crie uma nova VPC para você e seja usada para esse caso de uso. Nota: Só é relevante se VpcEnabledfor`Yes`.

IPAMPoolId	(Entrada opcional)	Se você quiser atribuir o intervalo CIDR usando o Amazon VPC IP Address Manager, forneça o ID do pool IPAM a ser usado. Nota: Só é relevante se VpcEnabledé `Yes` e CreateNewVpcéNo.
ExistingVpcId	(Entrada opcional)	ID da VPC de uma VPC existente a ser usada para o caso de uso. Nota: Só é relevante se VpcEnabledé `Yes` e CreateNewVpcéNo.
ExistingPrivateSubnetIds	(Entrada opcional)	Lista separada por vírgula IDs de sub-redes privadas existentes a serem usadas para implantar a função Lambda. Nota: Só é relevante se VpcEnabledé `Yes` e CreateNewVpcéNo.
ExistingSecurityGroupIds	(Entrada opcional)	Lista separada por vírgulas de grupos de segurança da VPC existente a ser usada para configurar funções Lambda. Nota: Só é relevante se VpcEnabledé `Yes` e CreateNewVpcéNo.

VpcAzs	(Entrada opcional)	Lista separada por vírgula AZs de onde as sub-redes do são criadas VPCs Nota: Só é relevante se VpcEnabledé `Yes` e CreateNewVpcéNo.
UseInferenceProfile	No	Se o modelo configurado for Bedrock, você poderá indicar se está usando o Bedrock Inference Profile. Isso garantirá que as políticas de IAM necessárias sejam configuradas durante a implantação da pilha. Para obter mais detalhes, consulte o seguinte https://docs.aws.amazon.com/bedrock/ latest/ userguide/cross -region-inference.html
Implantar UI	Sim	Selecione a opção para implantar a interface de usuário de front-end para essa implantação. Selecionar Não criará apenas a infraestrutura para hospedar o APIs, a autenticação do e o APIs processamento de back-end.

6. Escolha Avançar.
7. Na página Configurar opções de pilha, selecione Avançar.
8. Na página Revisar, verifique e confirme as configurações. Selecione a caixa confirmando que o modelo criará recursos do AWS Identity and Access Management (IAM).
9. Selecione Create stack (Criar pilha) para implantar a pilha.

Você pode ver o status da pilha no CloudFormation console da AWS na coluna Status. Você deve receber o status CREATE_COMPLETE em aproximadamente 10 a 30 minutos.

Implantação de um caso de uso autônomo do Bedrock Agent

Siga as step-by-step instruções nesta seção para configurar e implantar a solução em sua conta.

Tempo de implantação: aproximadamente 10 a 30 minutos

1. Faça login no AWS Management Console e selecione o botão para iniciar o CloudFront modelo.

BedrockAgent.modelo	

2. Por padrão, esse modelo é iniciado na região Leste dos EUA (Norte da Virgínia). Para iniciar a solução em outra região da AWS, use o seletor de Região na barra de navegação do console.

> **Note**
>
> Essa solução usa o Amazon Bedrock, que atualmente não está disponível em todas as regiões da AWS. Se você estiver usando esses recursos, deverá iniciar essa solução em uma região da AWS onde esses serviços estejam disponíveis. Para obter a disponibilidade mais atual por região, consulte a Lista de serviços regionais da AWS.

3. Na página Criar pilha, verifique se o URL de modelo correto é apresentado na caixa de texto URL do Amazon S3 e escolha Avançar.
4. Na página Especificar detalhes da pilha, atribua um nome para a sua pilha de soluções. Para obter informações sobre limitações de nomenclatura de caracteres, consulte {https---docs-aws-amazon-com- https---docs-aws-amazon-com -IAM-Latest- UserGuide -reference-iam-limits-html} [cotas do IAM e do AWS STS] no Guia do usuário do AWS Identity and Access Management.
5. Em Parâmetros, revise os parâmetros do modelo dessa solução e modifique-os conforme requerido. Esta solução usa os seguintes valores padrão.

Parâmetro	Entrada padrão	Description
UseCaseUUID	*<_Requires input_>*	36 caracteres UUIDv4 para identificar esse caso de uso implantado em um aplicativo.
UseCaseConfigRecordKey	*<Requires input>*	Chave correspondente ao registro que contém as configurações exigidas pela função Lambda do provedor de bate-papo em tempo de execução. O registro na tabela deve ter um atributo-chave correspon dente a esse valor e um atributo de configuração contendo a configuração desejada. Esse registro será preenchido pela plataforma de implantaç ão se estiver em uso. Para implantações autônomas desse caso de uso, é necessária uma entrada criada manualmente na tabela UseCaseConfigTable Namedefinida em.
UseCaseConfigTableName	*<Requires input>`*	A pilha lerá a configuração do caso de uso da tabela fornecida aqui e usará a chave de registro definida em UseCaseConfigRecordKey.

Parâmetro	Entrada padrão	Description
DefaultUserEmail	`placeholder@example.com`	E-mail do usuário padrão para esse caso de uso. A solução cria um usuário do Amazon Cognito para esse e-mail acessar o caso de uso.

Parâmetro	Entrada padrão	Description
ExistingRestApiId	(Entrada opcional)	ID da API REST da API Gateway existente a ser usada. Se não for fornecida , uma nova API REST do API Gateway será criada. Normalmente fornecido durante a implantação a partir do painel de implantação. Observação: usar o APIs Existing pode ajudar a reduzir a duplicação de recursos e simplificar o gerenciamento de APIs quando você precisa implantar vários casos de uso autônomos. Ao fornecer o existente APIs para um caso de uso independente, você é responsável por garantir que a API seja configurada com as rotas necessárias com os modelos esperados. Uma rota /details pré-configurada necessária (busca detalhes do caso de uso durante o bate-papo) e, opcionalm ente, uma rota /feedback (se FeedbackEnabledestiver definida como para permitir a coleta de feedback `Yes` para respostas de bate-papo do LLM) deve ser configura da. Além disso ExistingA piRootResourceId,, ExistingC ognitoUserPoolIde também

Parâmetro	Entrada padrão	Description
		ExistingCognitoGroupPolicyTableNamedeve ser fornecido.
ExistingApiRootResourceId	(Entrada opcional)	ID de recurso raiz da API Gateway REST existente a ser usado. O ID do recurso raiz da API REST pode ser obtido no console da AWS selecionando o recurso raiz (/) na seção “Recursos ” da API. O ID do recurso será então exibido no painel de detalhes do recurso. Como alternativa, você pode executar uma chamada de descrição da API em sua API REST para encontrar o ID do recurso raiz.
FeedbackEnabled	No	Se definido como Não, a pilha de casos de uso implantada não terá acesso ao recurso de feedback.
CognitoDomainPrefix	(Entrada opcional)	Insira um valor se quiser fornecer um domínio para o cliente do grupo de usuários do Amazon Cognito. Se você não fornecer um valor, a solução gerará um.

Parâmetro	Entrada padrão	Description
ExistingCognitoUserPoolId	(Entrada opcional)	UserPoolIdde um grupo de usuários existente do Amazon Cognito com o qual você deseja autenticar esse caso de uso. OBSERVAÇÃO: você normalmente fornece essa ID ao implantar a partir do painel de implantação, mas pode omiti-la ao implantar essa pilha de casos de uso de forma independente.
ExistingCognitoUserPoolClient	(Entrada opcional)	Forneça um cliente de grupo de usuários (cliente de aplicativo) para usar um existente. Se você não fornecer um cliente de grupo de usuários, a solução cria um. Você só pode fornecer esse parâmetro se tiver fornecido um ExistingCognitoUserPoolId.

Parâmetro	Entrada padrão	Description
ExistingCognitoGroupPolicyTableName	(Entrada opcional)	Nome da tabela do DynamoDB contendo políticas de grupos de usuários. Isso é usado pelo autorizador personalizado na API do caso de uso. OBSERVAÇÃO: Normalmente, você fornece esse nome ao implantar a partir do painel de implantação, mas pode omiti-lo ao implantar essa pilha de casos de uso de forma independente.
VpcEnabled	No	Se os recursos das pilhas serão implantados em uma VPC.
CreateNewVpc	No	Selecione Yes se quiser que a solução crie uma nova VPC para você e a use para esse caso de uso. NOTA: Esse parâmetro só é relevante se VpcEnabledforYes.
IPAMPoolId	(Entrada opcional)	Se você quiser atribuir o intervalo CIDR usando IPAM, forneça o ID do pool IPAM a ser usado. NOTA: Esse parâmetro só é relevante se VpcEnabledfor Yes e CreateNewVpcforNo.

Parâmetro	Entrada padrão	Description
ExistingVpcId	(Entrada opcional)	ID da VPC de uma VPC existente a ser usada para o caso de uso. NOTA: Esse parâmetro só é relevante se VpcEnabledfor `Yes` e CreateNewVpcforNo.
ExistingPrivateSubnetIds	(Entrada opcional)	Lista separada por vírgula IDs de sub-redes privadas existentes a serem usadas para implantar a função Lambda. NOTA: Esse parâmetro só é relevante se VpcEnabledfor `Yes` e CreateNewVpcforNo.
ExistingSecurityGroupIds	(Entrada opcional)	Lista separada por vírgulas de grupos de segurança da VPC existente a ser usada para configurar funções Lambda. NOTA: Esse parâmetro só é relevante se VpcEnabledfor `Yes` e CreateNewVpcforNo.
VpcAzs	(Entrada opcional)	Lista separada por vírgula AZs de onde as sub-redes do são criadas VPCs Nota: Só é relevante se VpcEnabledé `Yes` e CreateNewVpcéNo.
BedrockAgentId	*`<Requires input>`*	O ID do Amazon Bedrock Agent a ser usado.

Parâmetro	Entrada padrão	Description
BedrockAgentAliasId	*<Requires input>*	O ID de alias do Amazon Bedrock Agent a ser usado.
Implantar UI	`Yes`	Selecione a opção para implantar a interface de usuário de bate-papo de front-end para essa implantação. A seleção No resulta na criação da infraestrutura para hospedar o APIs, na autenticação do e no APIs processamento de back-end sem a interface do usuário do chat.

6. Escolha Avançar.
7. Na página Configurar opções de pilha, selecione Avançar.
8. Na página Revisar, verifique e confirme as configurações. Marque a caixa de seleção confirmando que o modelo criará recursos do IAM.
9. Selecione Create stack (Criar pilha) para implantar a pilha.

Você pode ver o status da pilha no CloudFormation console da AWS na coluna Status. Você deve receber o status CREATE_COMPLETE em aproximadamente 10 a 30 minutos.

Fornecendo uma configuração de chat do DynamoDB

Ao implantar um caso de uso, UseCaseConfigRecordKeyUseCaseConfigTableNamesão necessários CloudFormation parâmetros que normalmente são preenchidos pelo painel de implantação. A pilha de painéis de implantação gerencia a criação e a configuração dessa tabela, enquanto as chamadas para a API de implantação acionam o preenchimento dos parâmetros.

Ao realizar uma implantação autônoma, você deve fazer o seguinte:

1. Crie uma tabela do DynamoDB com uma chave de hash de chave.

2. Crie um registro na tabela contendo a configuração do caso de uso como um registro do formato: `{key: some_use_case_key, config: {your_configuration}.`
3. Passe os parâmetros escolhidos UseCaseConfigTableNamee UseCaseConfigRecordKey(`some_use_case_key`neste exemplo) para a pilha de casos de uso durante a implantação.

Para criar uma configuração adequada para uma implantação autônoma, você pode criar um caso de uso necessário no painel de implantação e copiar o registro da tabela de configuração. Caso contrário, você pode criar sua própria configuração com base no exemplo a seguir para uma implantação do Bedrock:

```
{
 "UseCaseName": "SampleUseCase",
 "ConversationMemoryParams": {
 "ConversationMemoryType": "DynamoDB",
 "HumanPrefix": "H",
 "AiPrefix": "A",
 "ChatHistoryLength": 20
 },
 "KnowledgeBaseParams": {
 "KnowledgeBaseType": "Bedrock",
 "NumberOfDocs": 2,
 "ScoreThreshold": 0,
 "ReturnSourceDocs": false,
 "BedrockKnowledgeBaseParams": {
 "BedrockKnowledgeBaseId": "SOME_ID",
 "OverrideSearchType": null
 }
 },
 "LlmParams": {
 "ModelProvider": "Bedrock",
 "BedrockLlmParams": { "ModelId": "anthropic.claude-v2" },
 "PromptParams": {
 "PromptTemplate": "some prompt",
 "MaxPromptTemplateLength": 187500,
 "MaxInputTextLength": 187500,
 "UserPromptEditingEnabled": true,
 "DisambiguationEnabled": true,
 "DisambiguationPromptTemplate": "some prompt"
 },
 "ModelParams": {},
```

```
 "Temperature": 1,
 "RAGEnabled": true,
 "Streaming": true,
 "Verbose": false
 }
}
```

Monitore a solução com o Service Catalog AppRegistry

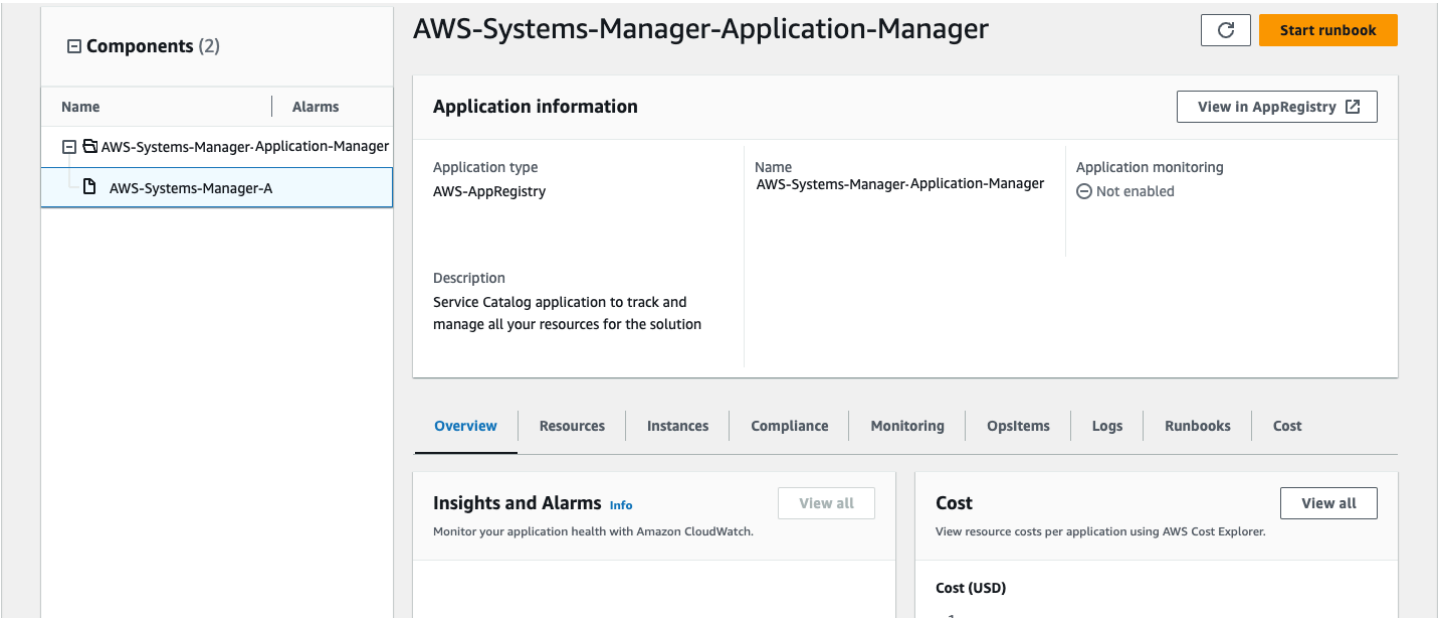
A solução inclui um AppRegistry recurso do Service Catalog para registrar o CloudFormation modelo e os recursos subjacentes como um aplicativo no Service Catalog AppRegistry e no Systems Manager Application Manager.

O Systems Manager Application Manager oferece uma visão em nível de aplicativo dessa solução e de seus recursos para que você possa:

- Monitore seus recursos, custos dos recursos implantados em pilhas, contas da AWS e registros associados a essa solução a partir de um local central.
- Visualize os dados operacionais dos recursos dessa solução no contexto de um aplicativo. Por exemplo, status de implantação, CloudWatch alarmes, configurações de recursos e problemas operacionais.

A figura a seguir mostra um exemplo da visualização do aplicativo para a pilha de soluções no Application Manager.

Descreve a pilha de soluções no Application Manager



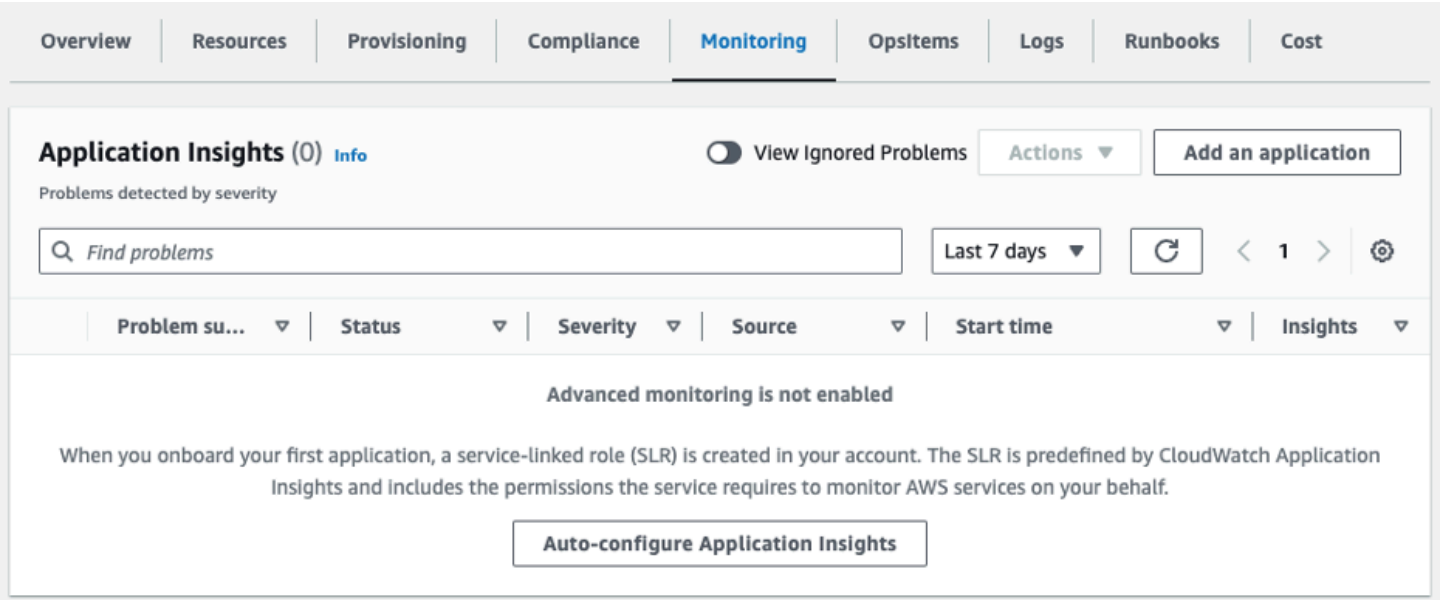
Ative CloudWatch Application Insights

1. Faça login no console do Systems Manager.

2. No painel de navegação, escolha Application Manager.
3. Em Aplicativos, pesquise o nome do aplicativo para essa solução e selecione-o.

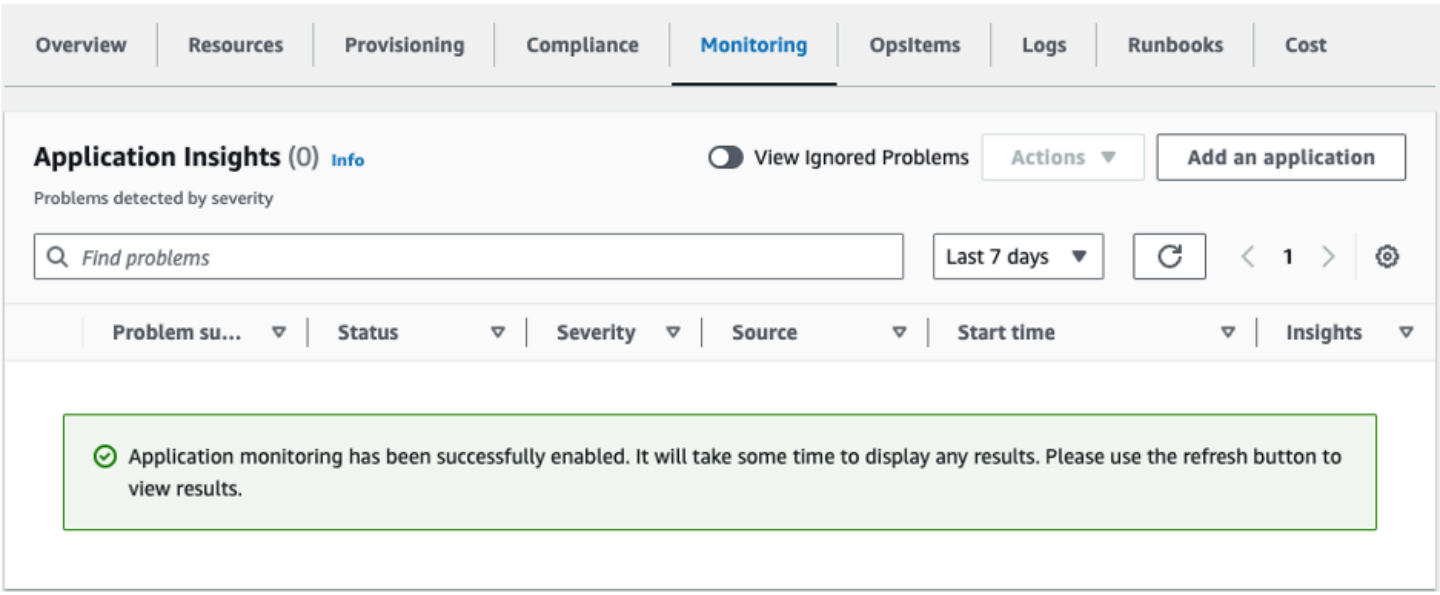
   O nome do aplicativo terá Registro do aplicativo na coluna Fonte do aplicativo e terá uma combinação do nome da solução, região, ID da conta ou nome da pilha.
4. Na árvore Componentes, escolha a pilha de aplicativos que você deseja ativar.
5. Na guia Monitoramento, em Application Insights, selecione Configurar automaticamente o Application Insights.

O painel do Application Insights não mostra problemas detectados e a opção de configuração automática.



O monitoramento de seus aplicativos agora está ativado e a seguinte caixa de status é exibida:

Painel do Application Insights mostrando a mensagem de ativação bem-sucedida do monitoramento.



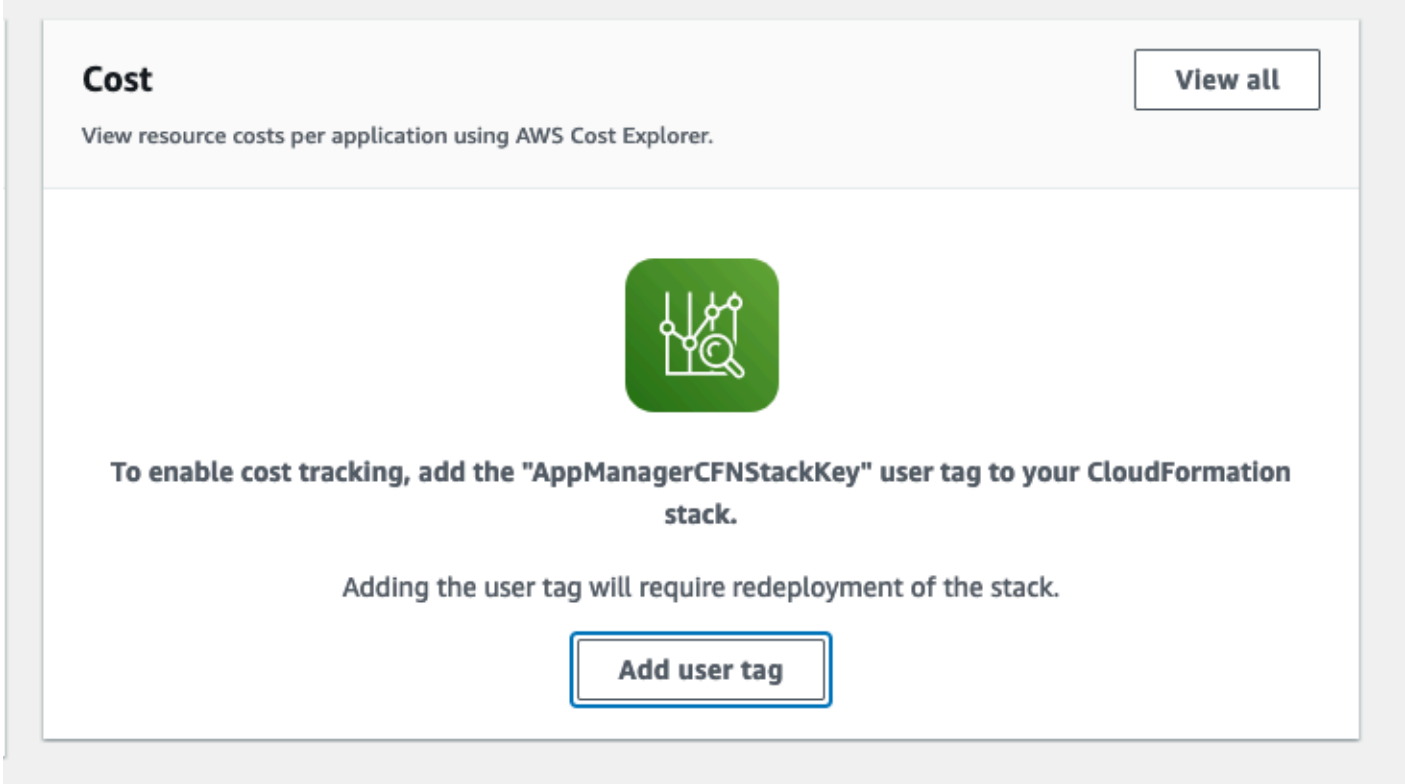
Confirme as tags de custos associadas à solução

Depois de ativar as etiquetas de alocação de custos associadas à solução, você deve confirmar as etiquetas de alocação de custos para ver os custos dessa solução. Para confirmar as tags de alocação de custos:

1. Faça login no console do Systems Manager.
2. No painel de navegação, escolha Application Manager.
3. Em Aplicativos, escolha o nome do aplicativo para essa solução e selecione-o.

   O nome do aplicativo terá Registro do aplicativo na coluna Fonte do aplicativo e terá uma combinação do nome da solução, região, ID da conta ou nome da pilha.
4. Na guia Visão geral, em Custo, selecione Adicionar tag de usuário.

   Captura de tela mostrando a tela de adição da tag de usuário do custo do aplicativo



5. Na página Adicionar tag de usuário, insira `confirm` e selecione Adicionar tag de usuário.

O processo de ativação pode levar até 24 horas para que os dados da tag apareçam.

Ative as tags de alocação de custos associadas à solução

Depois de ativar o Cost Explorer, ative as tags de alocação de custos associadas a essa solução para ver os custos dessa solução. As tags de alocação de custos só podem ser ativadas pela conta de gerenciamento da organização. Para ativar as tags de alocação de custos:

1. Faça login no console AWS Billing and Cost Management e Cost Management.
2. No painel de navegação, selecione Tags de alocação de custos.
3. Na página Tags de alocação de custos, filtre a tag AppManager CFNStack Key e selecione a tag nos resultados mostrados.
4. Selecione Ativar.

AWS Cost Explorer

Você pode ver a visão geral dos custos associados ao aplicativo e aos componentes do aplicativo no console do Application Manager por meio da integração com o AWS Cost Explorer, que deve ser ativado primeiro. O Cost Explorer ajuda você a gerenciar custos fornecendo uma visão dos custos e do uso dos recursos da AWS ao longo do tempo. Ativar o Cost Explorer para a solução:

1. Faça login no console de gerenciamento de custos da AWS.
2. No painel de navegação, selecione Cost Explorer para visualizar os custos e o uso da solução ao longo do tempo.

Atualizar a solução

Se você já implantou a solução, siga este procedimento para atualizar a CloudFormation pilha da solução e obter os recursos e aprimoramentos mais recentes. O processo de atualização tem três partes:

- Etapa 1: Atualizar o painel de implantação
- Etapa 2: migrar as configurações do caso de uso
- Etapa 3: atualizar casos de uso

> **Note**
>
> 1. Na versão 2.0.0, a integração com Anthropic e Hugging Face foi descontinuada em favor do Amazon Bedrock e do Amazon AI. SageMaker Você pode implantar modelos disponíveis por meio do Hugging Face. SageMaker JumpStart Consulte Use Hugging Face with Amazon SageMaker AI para obter mais detalhes.
> 2. Certifique-se de testar o processo de atualização em um ambiente que não seja de produção antes de executar essas etapas.

Etapa 1: Atualizar o painel de implantação

1. Entre no CloudFormation console, selecione sua CloudFormation pilha existente e selecione Atualizar.
2. Selecione Substituir modelo atual.
3. Em Especificar modelo:
   a. Selecione URL do Amazon S3.
   b. Copie o link do CloudFormation modelo mais recente.
   c. Cole o link na caixa de URL do Amazon S3.
   d. Verifique se o URL do modelo correto aparece na caixa de texto URL do Amazon S3 e escolha Avançar. Escolha Avançar novamente.
4. Em Parâmetros, revise os parâmetros do modelo e modifique-os conforme necessário. Para obter detalhes sobre os parâmetros, consulte Etapa 1: Iniciar a pilha do painel de implantação.

5. Escolha Avançar.
6. Na página Configurar opções de pilha, selecione Avançar.
7. Na página Revisar, verifique e confirme as configurações. Marque a caixa de seleção para confirmar que o modelo cria recursos do IAM.
8. Escolha Exibir conjunto de alterações e verifique as alterações.
9. Selecione Criar pilha para implantar a pilha.

Você pode ver o status da pilha no CloudFormation console da AWS na coluna Status. Você deve receber o status UPDATE_COMPLETE em aproximadamente 10 minutos.

Se a versão existente da Solução for anterior à v2.0.0, a atualização criará uma pilha de interface do usuário da web (que substitui a `amplify-ui` implementação da tela de login por uma interface de usuário hospedada no Cognito) e uma nova CloudFront URL, que pode ser obtida na seção Saída do CloudFormation console quando o status da pilha for UPDATE_COMPLETE.

> **Note**
>
> Os casos de uso existentes criados usando versões anteriores à v2.0.0 NÃO serão exibidos até que você conclua as etapas descritas abaixo.

Etapa 2: migrar configurações de casos de uso (somente atualizações de versões anteriores à 2.0.0)

O esquema para armazenamento e o serviço da AWS para armazenar a configuração do caso de uso foram alterados na versão 2.0.0. Siga as etapas descritas no Guia do usuário de migração do GAAB v2 usando o script gaab_v2_migration.py. Depois de executar o script, você pode acessar o painel de implantação para ver os casos de uso implantados.

> **Note**
>
> Você deve seguir as etapas abaixo para concluir a migração dos casos de uso.

Etapa 3: atualizar casos de uso

Você pode editar os casos de uso implantados com os novos recursos disponíveis nas versões mais recentes do GAAB. Consulte Usar a solução para obter informações sobre como usar os recursos dessa solução.

Para atualizar os casos de uso para a versão mais recente, você deve concluir as etapas de `Editar`caso de uso no painel de implantação (embora você possa não fazer nenhuma alteração). Essa ação aciona uma atualização da CloudFormation pilha com a versão mais recente do modelo.

> **Note**
>
> Casos de uso criados com versões 1.x ou 2.x da solução podem não funcionar com versões posteriores. Portanto, recomendamos clonar casos de uso existentes criados com versões anteriores à v3.0.0 por meio do painel de implantação. Em seguida, migre gradualmente e substitua por novos casos de uso criados usando a versão 3.0.0 ou posterior.

Solução de problemas

Esta seção fornece instruções de solução de problemas para implantar e usar a solução.

Se essas instruções não resolverem seu problema, Entre em contato com o suporte fornecerá instruções para abrir um caso de suporte para essa solução.

Problema: a implantação de uma configuração habilitada para VPC, com Create a VPC for me, falha

A pilha do painel de implantação ou a pilha de casos de uso falha na implantação porque não CloudFormation foi possível provisionar recursos de rede VPC.

Resolução

Verifique os limites de cota e a VPCs Elastic IPs em sua conta. Os limites padrão são 5 para cada Elastic IPs e VPCs por conta da AWS, por região da AWS.

> **Note**
>
> Quando a solução cria uma VPC, uma única implantação habilitada para VPC (painel de implantação ou caso de uso) é uma implantação de 2 AZ com 1 sub-rede pública e 1 sub-rede privada em cada AZ, cada sub-rede pública implanta 1 gateway NAT. Com 2 gateways NAT, a implantação consome 2 endereços IP públicos do limite de cota.

Alguns limites que você deve conhecer (por conta, por região):

- Número de VPCs - 5
- Número de endereços IP públicos - 5
- Número de endpoints VPC do gateway - 20
- Número de endpoints VPC de interface: 20

Problema: a pilha de casos de uso não pode ser excluída CloudFormation após a exclusão da pilha do painel de implantação

Se a pilha do painel de implantação for excluída CloudFormation antes que todas as pilhas de casos de uso sejam excluídas, os casos de uso podem acabar em um estado bloqueado (inutilizável). Isso ocorre porque uma função do IAM criada pela pilha do painel de implantação não existe mais, impedindo modificações na pilha de casos de uso.

Resolução

> **Warning**
>
> Certifique-se de limpar todas as funções criadas manualmente imediatamente após o uso. Essas são permissões elevadas que os usuários podem explorar para elevar a função.

Recrie a função excluída do IAM para permitir a exclusão das CloudFormation pilhas:

1. Abra o CloudFormation console e determine a função associada à sua pilha bloqueada.
   a. O ARN da função pode ser encontrado na seção de informações da pilha chamada função do IAM.
   b. O nome da função é o que vem depois de:role/ no ARN da função do IAM (por exemplo, arn:aws:iam: ::role/) <account-id><role-name>
2. Crie uma nova função no IAM com o mesmo nome da função excluída.
   a. Selecione o serviço da AWS como a entidade confiável e CloudFormationselecione no menu suspenso.
   b. Adicione as permissões necessárias. Se não tiver certeza sobre as permissões necessárias, você pode usar a AdministratorAccesspolítica gerenciada da AWS.
   c. Insira o nome da função exatamente como obtido na Etapa 1.
3. Volte ao CloudFormation console e exclua as pilhas bloqueadas.
4. Depois que todas as pilhas bloqueadas forem excluídas com sucesso, retorne ao IAM e exclua todas as funções criadas na Etapa 2.

Problema: a interface do usuário do caso de uso não reflete as alterações nas configurações

Quando os casos de uso são atualizados, a interface do usuário é implantada em. CloudFront No entanto, como armazena em CloudFront cache as implantações e o arquivo de configuração que determina como algumas configurações são mostradas ao usuário, essas alterações podem não ser refletidas imediatamente.

Resolução

A CloudFront distribuição pode ser invalidada para forçar a propagação da nova configuração para usuários front-end.

1. Abra o CloudFormation console e determine a CloudFront distribuição associada à sua pilha de casos de uso.
   a. A pilha de casos de uso deve começar com o mesmo nome que você usou ao implantar o caso de uso.
   b. Localize a pilha aninhada correspondente à interface do usuário. O nome da pilha aninhada deve começar com WebAppS3 UINested StackS3. UINested StackResource
   c. Na guia Recursos, localize o tipo AWS::CloudFront::Distributionde recurso e selecione a ID física. Isso abrirá a distribuição no CloudFront console.
2. Navegue até a guia Invalidações, escolha Criar invalidação e insira um caminho de /*. Isso invalidará todos os caminhos.
3. Em seu próprio navegador, exclua todos os cookies e arquivos em cache relacionados ao caso de uso.

Entrar em contato com o AWS Support

Se você tem o AWS Business Support+, o AWS Enterprise Support ou o Unified Operations, você pode usar o AWS Support Center para obter assistência especializada com essa solução. As seções a seguir dão instruções.

Criar caso

1. Faça login na Central de suporte.
2. Escolha Criar caso.

Como podemos ajudar?

1. Escolha Técnico.
2. Em Serviço, selecione Soluções.
3. Em Categoria, selecione Outras soluções.
4. Em Severidade, selecione a opção que melhor corresponda ao seu caso de uso.
5. Quando você insere o Serviço, a Categoria e a Gravidade, a interface preenche links para perguntas comuns de solução de problemas. Se você não conseguir resolver sua pergunta com esses links, escolha Próxima etapa: mais informações.

Mais informações

1. Em Assunto, insira um texto resumindo sua pergunta ou problema.
2. Para obter uma descrição, descreva o problema em detalhes, incluindo o nome dessa solução: Generative AI Application Builder na AWS.
3. Selecione Anexar arquivos.
4. Anexe as informações que o AWS Support precisa para processar a solicitação.

Ajude-nos a resolver seu caso com mais rapidez

1. Insira as informações solicitadas.
2. Escolha Próxima etapa: solucione ou entre em contato conosco.

Solucione ou entre em contato conosco

1. Analise as soluções Solucionar agora.
2. Se você não conseguir resolver seu problema com essas soluções, escolha Fale conosco, insira as informações solicitadas e escolha Enviar.

Desinstalar a solução

> **Note**
>
> As implantações criadas por meio do painel de implantação não devem ser gerenciadas fora da solução. Certifique-se de excluir e limpar todas as implantações no painel de implantação antes de excluir a pilha. CloudFormation

Você pode desinstalar o Generative AI Application Builder na solução AWS a partir do AWS Management Console ou usando a AWS Command Line Interface. Você deve excluir manualmente os buckets do Amazon S3, os índices do Amazon Kendra ou os registros criados por essa solução. CloudWatch As soluções da AWS não excluem automaticamente buckets do Amazon S3, índices do Amazon Kendra ou CloudWatch registros, caso você tenha armazenado dados para reter.

Como usar o AWS Management Console

1. Faça login no CloudFormation console da AWS.
2. Na página Pilhas, selecione a pilha de instalação dessa solução.
3. Escolha Excluir.

Usar a AWS Command Line Interface

Determine se a AWS Command Line Interface (AWS CLI) está disponível em seu ambiente. Para obter instruções de instalação, consulte O que é a interface de linha de comando da AWS no Guia do usuário da AWS CLI. Depois de confirmar que a AWS CLI está disponível, execute o seguinte comando:

```
$ aws cloudformation delete-stack --stack-name <installation-stack-name>
```

Etapas de desinstalação manual

Excluindo os buckets do Amazon S3

Essa solução está configurada para reter o bucket Amazon S3 criado pela solução se você decidir excluir a CloudFormation pilha da AWS para evitar perda acidental de dados. Depois de desinstalar a solução, você pode excluir manualmente esse bucket do Amazon S3 se não precisar reter os dados. Siga estas etapas para excluir o bucket do Amazon S3.

1. Faça login no console do Amazon S3.
2. No painel de navegação, selecione Buckets.
3. Localize os <stack-name>buckets do S3.
4. Selecione o bucket do S3 e escolha Excluir.

Para excluir o bucket do S3 usando o AWS CLI, execute o comando a seguir. Você não precisará esvaziar o bucket primeiro ao usar a opção --force.

```
$ aws s3 rb s3://<bucket-name> --force
```

Excluindo os índices do Amazon Kendra

Para evitar a perda acidental de dados, essa solução é configurada para reter os índices do Amazon Kendra criados pela solução quando a pilha da AWS for excluída. CloudFormation Depois de desinstalar a solução, você pode excluir manualmente os índices do Amazon Kendra para os quais não precisa mais reter dados. Siga estas etapas para excluir o índice Amazon Kendra.

1. Faça login no console do Amazon Kendra.
2. No painel de navegação, selecione Índices.
3. Localize e selecione o índice que você deseja excluir.
4. Selecione Excluir para excluir o índice selecionada.

Para excluir o índice do Amazon Kendra usando o AWS CLI, execute o seguinte comando:

```
$ aws kendra delete-index --id<index-id>
```

Excluindo os registros CloudWatch

Para evitar a perda acidental de dados, configuramos essa solução para reter os CloudWatch registros caso você decida excluir a CloudFormation pilha. Depois de desinstalar a solução, você pode excluir manualmente os registros se não precisar reter os dados. Siga estas etapas para excluir os CloudWatch registros.

1. Faça login no CloudWatch console da Amazon.
2. No painel de navegação, selecione Grupos de registros.
3. Localize os grupos de registros criados pela solução.
4. Selecione um dos grupos de registros.
5. Escolha Ações e, em seguida, escolha Excluir.

Repita as etapas até excluir todos os grupos de registros de soluções.

Uso da solução

Acessando a interface do usuário

Durante o processo de implantação da pilha (tanto para o painel de implantação quanto para os casos de uso), um e-mail é enviado para o endereço de e-mail configurado. O e-mail contém as credenciais temporárias do usuário que ele pode usar para se inscrever e acessar a interface da web.

> **Note**
>
> O DevOps usuário com acesso ao AWS Management Console deve fornecer ao usuário administrador a CloudFront URL da interface do usuário do painel de implantação quando a pilha for concluída.

Para os casos de uso, o usuário administrador com acesso à interface do usuário do painel de implantação deve fornecer ao usuário comercial a CloudFront URL da interface do usuário do caso de uso quando a implantação for concluída.

Uma vez logado, o usuário pode interagir com a solução UIs, seja no painel de implantação, no caso de administradores, ou no caso de uso, no caso de usuários corporativos.

Como atualizar uma implantação

Quando estiver na página inicial do painel de implantação (ou na página de detalhes de uma implantação), você pode editar a configuração usada por uma implantação. Você só pode editar implantações que estejam nos status CREATE_COMPLETE ou UPDATE_COMPLETE.

Com exceção do nome do caso de uso, todas as outras opções são editáveis para uma implantação. Basta alterar os valores que você deseja editar e reimplantar.

Dependendo do escopo das edições feitas, o tempo de redistribuição variará. Pode levar alguns segundos se configurações simples tiverem sido alteradas (por exemplo, parâmetros do modelo), até mais de 30 minutos se opções maiores relacionadas à infraestrutura tiverem sido alteradas (exemplo, solicitação para criar o índice Amazon Kendra para o caso de uso de texto RAG).

Depois que a edição for concluída com sucesso, o status do aplicativo relatará um status UPDATE_COMPLETE. No momento, você pode acessar a interface implantada por meio da CloudFront URL e interagir com a implantação modificada.

> **Note**
>
> Talvez seja mais fácil executar várias implantações side-by-side se você quiser comparar configurações diferentes ou LLMs. Use o recurso Clone para usar rapidamente uma configuração existente para iniciar uma nova implantação.

Como clonar uma implantação

Quando estiver na página inicial do painel de Implantações (ou na página de detalhes de uma implantação), você pode clonar a configuração usada por uma implantação. A clonagem de uma implantação inicia o assistente Deploy new use case, mas com a maioria dos campos pré-preenchidos com os mesmos valores.

Essa é uma operação conveniente para ajudá-lo a duplicar rapidamente implantações com configurações alteradas, reviver uma implantação excluída ou comparar várias implantações LLMs idênticas.

Como excluir uma implantação

Quando estiver na página inicial do painel de Implantações (ou na página de detalhes de uma implantação), você poderá excluí-la quando não precisar mais da implantação. A exclusão de uma implantação invoca uma operação de exclusão de CloudFormation pilha e desprovisiona os recursos para a implantação.

Por padrão, uma implantação excluída ainda permanece no painel para ativar a funcionalidade de clonagem. Para remover completamente uma implantação do painel para que ela deixe de ser rastreada na interface do usuário, escolha Excluir permanentemente na janela de confirmação de exclusão.

> **Important**
>
> Alguns recursos são deixados para trás durante a exclusão da pilha e devem ser excluídos manualmente. Consulte a seção Desinstalação manual para obter detalhes sobre quais recursos são retidos e como limpá-los.

Configurando um modelo de linguagem grande (LLM)

Qual LLM é ideal para seu caso de uso depende de um grande conjunto de fatores específicos às suas necessidades e do tipo de experiência do cliente que você deseja organizar. Essa solução não parece ser prescritiva, mas tem como objetivo fornecer as ferramentas necessárias para avaliar o que funciona melhor para sua aplicação.

O espaço gerado pela IA está evoluindo rapidamente, portanto, cabe a você se manter atualizado sobre os modelos, as técnicas de otimização e as melhores práticas mais recentes para garantir que esteja criando as experiências certas para seus clientes.

> **Note**
>
> Se você estiver trabalhando com dados confidenciais ou não públicos, certifique-se de selecionar uma opção de LLM usando os serviços da AWS (como Amazon Bedrock ou Amazon SageMaker AI). Isso melhora a postura geral de segurança de sua implantação ao manter os dados em sua região e na rede AWS em comparação com o uso de um LLM hospedado por um provedor terceirizado.

Usando o Amazon SageMaker AI como um provedor de LLM

A partir da versão 1.3.0, o Amazon SageMaker AI está disponível como provedor de modelos para casos de uso de texto. Esse recurso permite que você use um endpoint de inferência de SageMaker IA já existente na conta da AWS na solução. Aqui estão algumas maneiras de começar.

> **Important**
>
> A solução não gerencia o ciclo de vida de seus endpoints de SageMaker IA. Você é responsável por excluir os endpoints de SageMaker IA quando eles não forem mais necessários para parar de incorrer em cobranças adicionais.

Criação de um endpoint de SageMaker IA

Você pode usar o Amazon SageMaker AI JumpStart para implantar rapidamente um endpoint.

Você também pode usar um endpoint de SageMaker IA baseado em geração de texto e implantar usando o serviço básico SageMaker de IA. Consulte a JumpStart documentação de SageMaker IA para obter um guia passo a passo sobre como implantar um modelo para inferência.

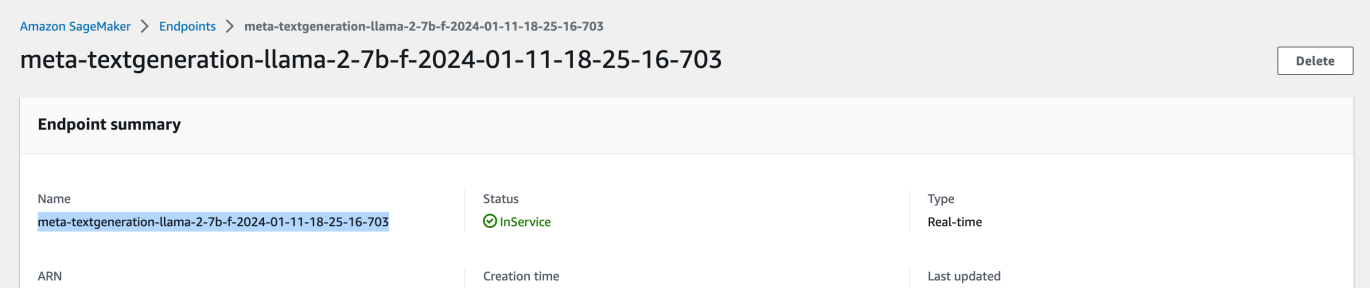
> **Note**
>
> models/LLMs As bases geralmente são muito grandes e geralmente podem exigir o uso de grandes instâncias de computação acelerada. Muitas dessas instâncias maiores podem não estar disponíveis por padrão em sua conta da AWS. Consulte as cotas de SageMaker IA padrão e certifique-se de solicitar um aumento de cota antes da implantação para evitar falhas comuns de implantação.

Use o endpoint de SageMaker IA para criar uma implantação de caso de uso de texto

Para implantar um novo caso de uso de texto usando um endpoint de SageMaker IA para inferência:

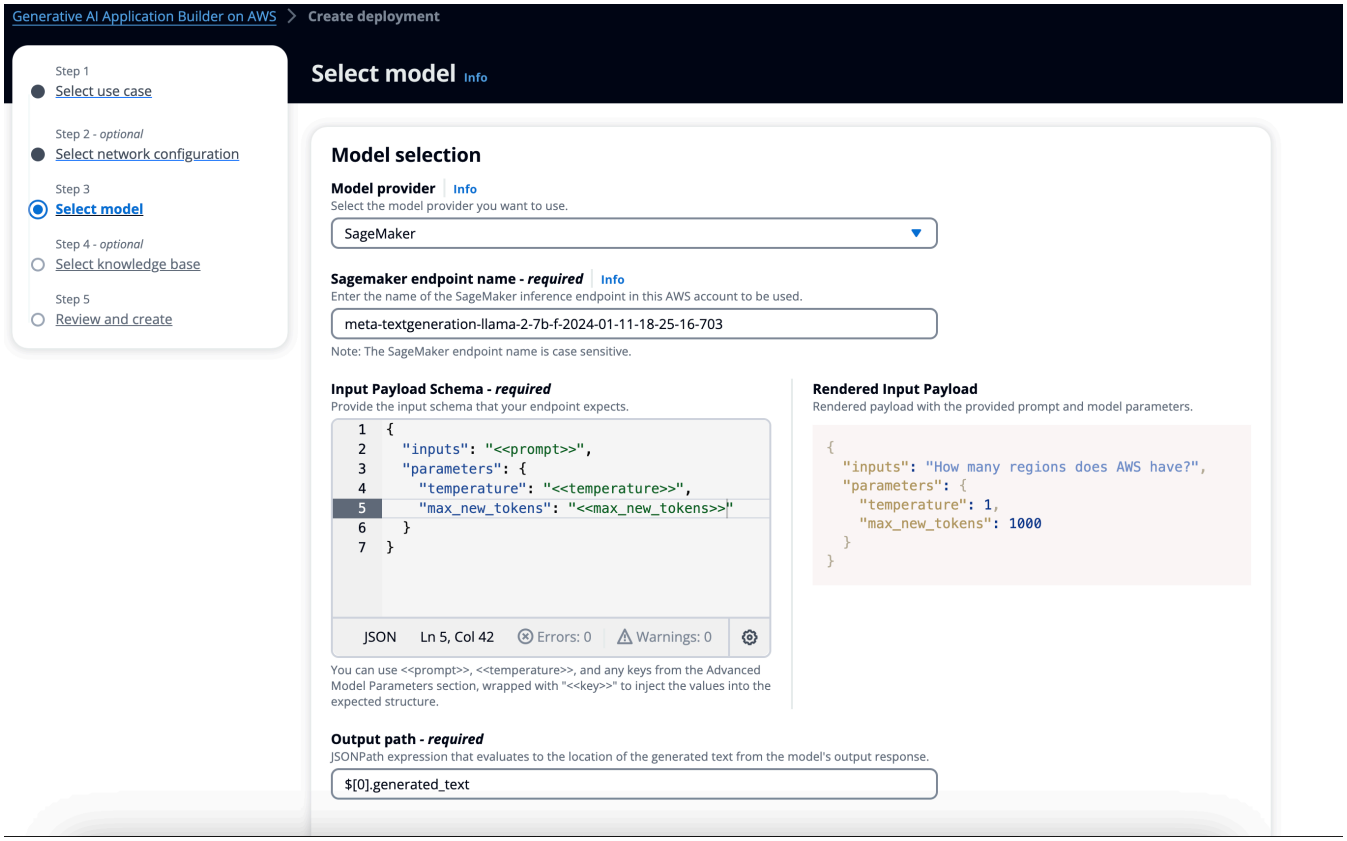
1. Crie um novo caso de uso por meio do assistente do painel de implantação e preencha os formulários até chegar à página de seleção de modelos.
2. Na página Modelos, selecione SageMaker AI como fornecedora de modelos. Isso gerará um formulário personalizado que exige três partes principais de entrada do usuário:
   - O nome do endpoint de SageMaker IA que você deseja usar. DevOps os usuários podem obter isso no console da AWS. Observe que o endpoint deve estar na mesma conta e região em que a solução está implantada.

     Localização do nome do endpoint no console da AWS



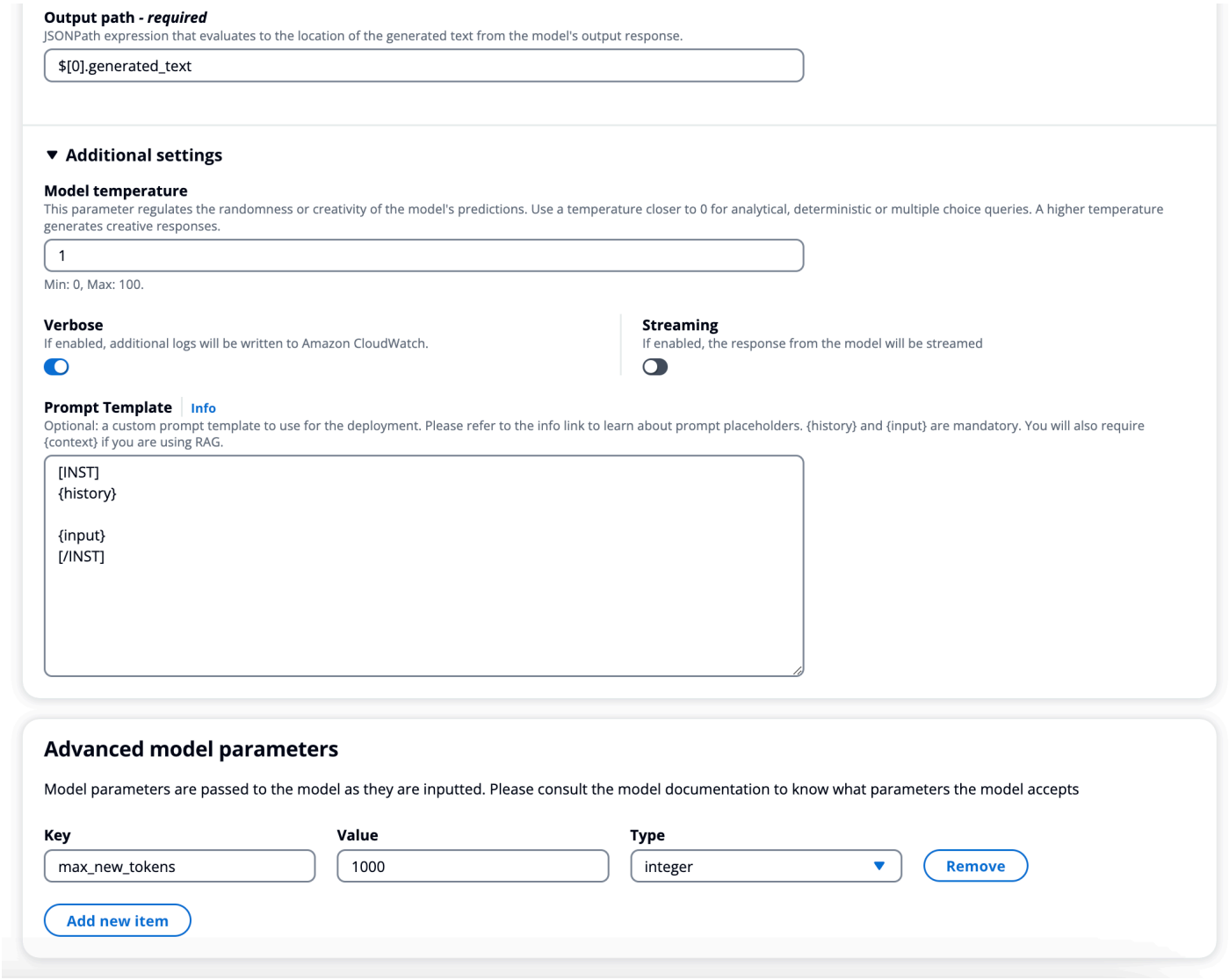
- O esquema da carga de entrada esperada pelo endpoint. Para oferecer suporte ao maior conjunto de endpoints, os usuários administradores precisam informar à solução como o endpoint espera que a entrada seja formatada. No assistente de seleção de modelo, forneça o esquema JSON para a solução ser enviada ao endpoint. Você pode adicionar espaços reservados para injetar valores estáticos e dinâmicos na carga útil da solicitação. As opções disponíveis são:
  - Espaços reservados obrigatórios:\ < <prompt\ >\\ > serão substituídos dinamicamente pela entrada completa (por exemplo, histórico, contexto e entrada do usuário de acordo com o modelo de prompt) a ser enviada ao endpoint de SageMaker IA em tempo de execução.
  - Espaços reservados opcionais:\ < <temperature\ >\\ > *,\ *, bem como quaisquer parâmetros definidos nos parâmetros avançados do modelo, podem ser fornecidos ao endpoint. Qualquer string contendo um espaço reservado entre\ <\ < and\ >\ > (por exemplo,\ <\ <max_new_tokens\ >\ >) será substituída pelo valor do parâmetro de modelo avançado com o mesmo nome.

    Exemplo de esquema de entrada - definindo campos obrigatórios, aviso e temperatura, junto com um parâmetro avançado personalizado, max_new_tokens. O caminho de saída deve ser fornecido como uma JSONPath string válida



3. A localização da resposta de string LLMs gerada na carga de saída. Isso deve ser fornecido como uma JSONPath expressão para indicar onde se espera que a resposta de texto final mostrada aos usuários seja acessada de dentro do objeto de retorno e da resposta do endpoint.

   Exemplo de adição de parâmetros de modelo avançados para uso no esquema de entrada de SageMaker IA (consulte a Figura 2 para ver as opções/configurações anteriores)



> **Note**
>
> SageMaker Agora, a IA oferece suporte à hospedagem de vários modelos no mesmo endpoint, e essa é a configuração padrão ao implantar um endpoint na versão atual do SageMaker AI Studio (não no Studio Classic).
> Se seu endpoint estiver configurado dessa forma, você deverá adicionar InferenceComponentNameà seção de parâmetros avançados do modelo, um valor correspondente ao nome do modelo que você deseja usar.

Configurações avançadas do LLM

Ao usar o Amazon Bedrock, você pode definir algumas configurações avançadas para seus modelos, como Amazon Bedrock Guardrails, Provisioned Throughput for Amazon Bedrock e parâmetros adicionais do modelo.

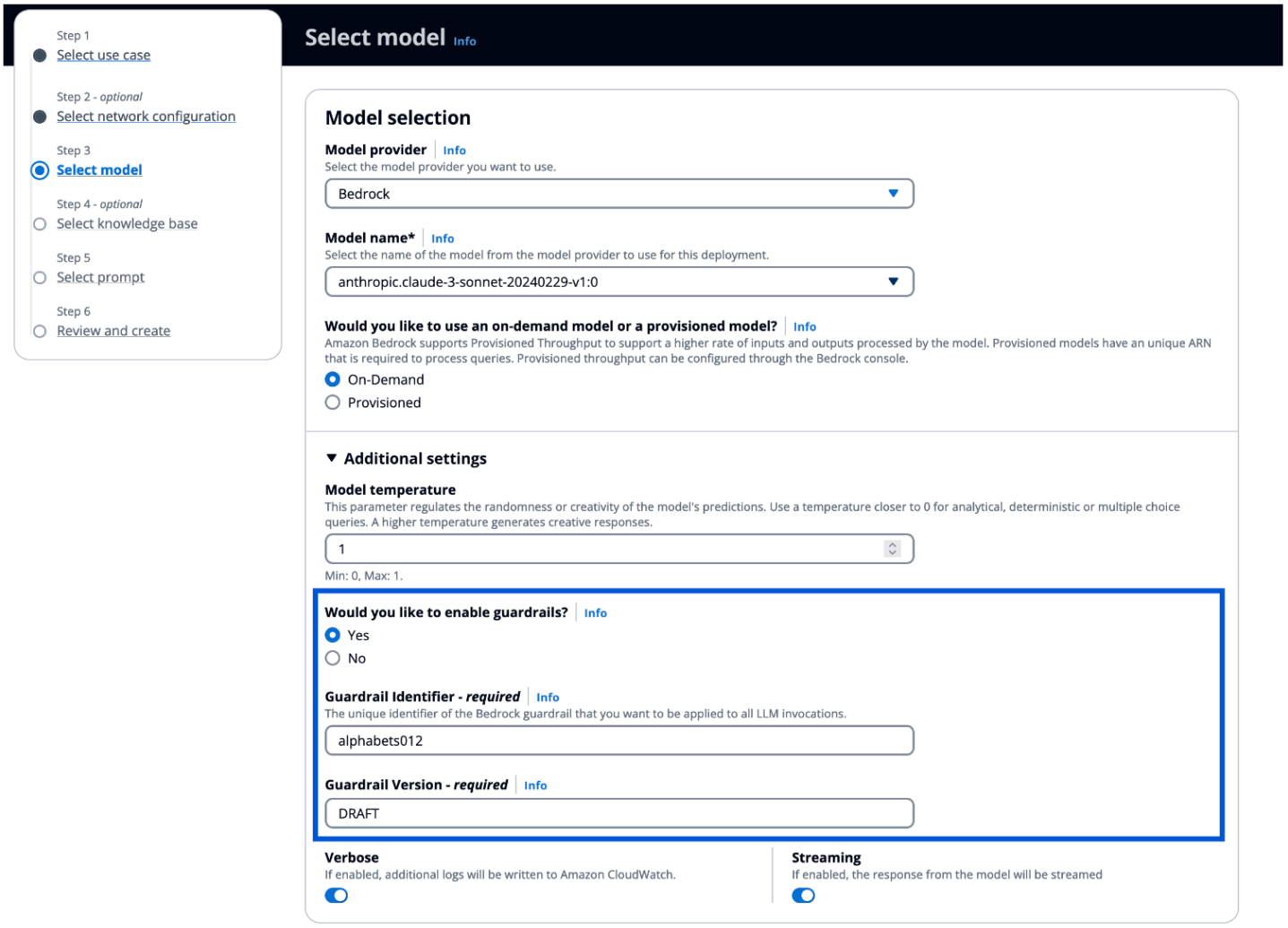
Barreiras de proteção do Amazon Bedrock

O Amazon Bedrock Guardrails é um recurso do Amazon Bedrock que avalia as entradas do usuário e as respostas do LLM com base nas políticas configuradas pelo usuário e fornece uma camada adicional de salvaguardas, independentemente do LLM subjacente que o usuário selecione para um caso de uso. Um Guardrail consiste em duas políticas para evitar conteúdo que se enquadre em categorias indesejáveis ou prejudiciais:

1. Tópicos negados para definir um conjunto de tópicos indesejáveis no contexto do aplicativo do usuário, por exemplo, consultoria de investimento em um aplicativo financeiro e,
2. Filtros de conteúdo****que permitem filtrar as solicitações de entrada do usuário ou as respostas do modelo contendo conteúdo nocivo.

Para uso na solução Generative AI Application Builder, um guardrail deve ser configurado no console Amazon Bedrock usando o assistente Create guardrail. Depois de criado, você pode adicionar esse Guardrail ao seu caso de uso de chat criado por meio do assistente de solução Generative AI Application Builder nas Configurações adicionais na etapa Seleção de modelo, fornecendo seu Identificador de Guardrail e sua versão de Guardrail.

Representa o assistente de implantação - habilitando o Amazon Bedrock Guardrails



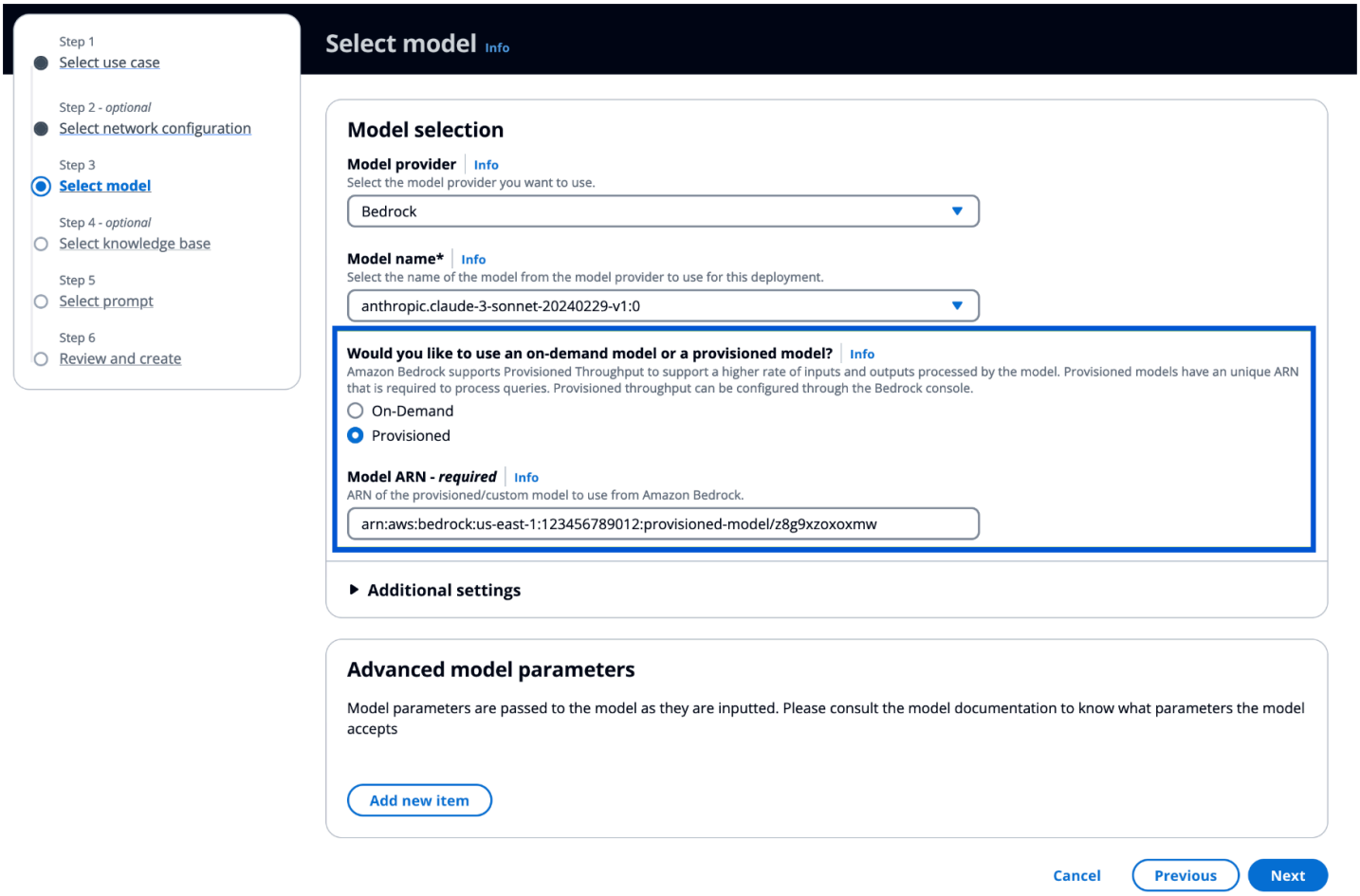
Taxa de transferência provisionada para Amazon Bedrock

Cada modelo sob demanda do Amazon Bedrock segue o limite de cota de conta específico da região para inferência de modelos. Por exemplo, o Anthropic Claude 2.x no Bedrock atualmente permite 500 solicitações e 500.000 tokens processados por minuto nas regiões us-east-1 e us-west-2. Talvez você também queira usar a solução com seus modelos pré-treinados ajustados ou contínuos. Para esses casos, o Amazon Bedrock permite uma taxa de transferência provisionada que permite executar grandes cargas de trabalho de inferência consistentes para seus modelos pré-treinados básicos, ajustados ou contínuos para uso em aplicativos de nível de produção.

Depois que a taxa de transferência provisionada é comprada no console Amazon Bedrock, um ARN de modelo é gerado para uso. Agora você pode fornecer esse ARN do modelo no assistente Generative AI Application Builder na etapa de seleção do modelo. Para fazer isso, selecione Bedrock como fornecedor do modelo e o nome do modelo base que foi usado para gerar esse ARN de

modelo provisionado no console Amazon Bedrock. Em seguida, selecione “Modelo provisionado” ao escolher entre modelos sob demanda e provisionados e forneça o ARN do modelo.

Descreve o assistente de implantação - Habilitando a taxa de transferência provisionada para o Amazon Bedrock



> **Note**
>
> Sua grade de proteção e taxa de transferência provisionada devem estar na mesma região do Painel de Implantação implantado e das pilhas de casos de uso.

Parâmetros do modelo

LLMs geralmente aceitam uma ampla gama de parâmetros específicos para sua implementação. Os fornecedores de modelos geralmente fornecem documentação descrevendo o conjunto de parâmetros suportados e seus usos.

A solução passa os parâmetros do modelo diretamente para o modelo subjacente, por isso é importante garantir que os parâmetros sejam definidos corretamente. Consulte a documentação do fornecedor do modelo para obter as informações mais recentes sobre os parâmetros suportados.

Configurando o Agent Builder

O Agent Builder fornece opções de configuração abrangentes para criar agentes de IA prontos para produção. Esta seção descreve como configurar e gerenciar as implantações do Agent Builder.

Configuração do prompt do sistema

O prompt do sistema define o comportamento, a personalidade e as capacidades do seu agente. Para configurar o prompt do sistema:

1. No assistente do Agent Builder, navegue até a etapa Configurar agente.
2. Edite o modelo de prompt do sistema no editor de texto.
3. Inclua instruções claras para:
   - Papel e propósito do agente
   - Como usar as ferramentas disponíveis (servidores MCP)
   - Preferências de formatação de resposta
   - Diretrizes comportamentais
4. Use o botão Redefinir para o padrão para restaurar o modelo original, se necessário.

Práticas recomendadas para solicitações de agentes:

- Seja específico sobre as capacidades e limitações do agente
- Forneça exemplos claros do comportamento desejado
- Inclua instruções para o uso da ferramenta e quando invocá-las
- Defina as expectativas do formato de resposta
- Estabeleça limites para o comportamento do agente

Integração de servidor MCP

Os servidores Model Context Protocol (MCP) fornecem aos agentes acesso a ferramentas corporativas e fontes de dados. Para configurar servidores MCP:

1. Na etapa Configurar Agente, localize a seção Servidores MCP.
2. Selecione entre os servidores MCP disponíveis no menu suspenso.

> **Note**
>
> Os servidores MCP devem estar configurados e acessíveis antes da implantação do agente. O agente descobrirá e usará automaticamente as ferramentas expostas pelos servidores MCP configurados. Consulte a documentação do MCP para obter informações sobre a configuração do servidor e da ferramenta.

Memory Settings

O Agent Builder fornece dois tipos de memória para manter o contexto e o conhecimento:

Memória de curto prazo

Ativado por padrão para todos os agentes:

- Mantém o contexto da conversa nas sessões
- Captura automaticamente as mensagens do usuário e as respostas do agente
- Organizado por ActorID e SessionID para isolamento adequado
- Nenhuma configuração é necessária

Memória de longo prazo

Recurso opcional para armazenar insights em todas as sessões:

1. Na etapa Configurar agente, localize a seção Configuração de memória.
2. Alterne a opção Ativar memória de longo prazo para ativar.
3. Quando ativado, o agente pode:
   - Extraia e armazene informações importantes em todas as conversas
   - Recupere o contexto relevante de sessões anteriores
   - Desenvolva conhecimento sobre as preferências e o histórico do usuário

> **Note**
>
> A memória de longo prazo usa AgentCore Memória com estratégia de memória semântica e configurações de retenção padrão.

Monitorando implantações do Agent Builder

O Agent Builder fornece monitoramento abrangente por meio de CloudWatch painéis e métricas.

Acessando CloudWatch painéis

1. Navegue até o CloudWatch console em sua conta da AWS.
2. Selecione Painéis na navegação à esquerda.
3. Encontre o painel chamado`AgentBuilder-<UseCaseId>`.
4. Visualize métricas em tempo real e dados históricos de desempenho.

Acesso e análise de registros

Os registros do agente estão disponíveis em CloudWatch Registros:

1. Navegue até CloudWatch Logs no console da AWS.
2. Encontre grupos de registros prefixados com`/aws/bedrock-agentcore/runtimes/`.
3. Use o CloudWatch Insights para consultar e analisar registros.
4. Pesquise padrões específicos de solicitação IDs ou erro.

Configurando o criador de fluxo de trabalho

O Workflow Builder permite a orquestração de vários agentes por meio de um agente supervisor que delega o trabalho a agentes especializados do Agent Builder.

Criação de um fluxo de trabalho

1. Navegue até o Painel de Implantação
2. Selecione Criar caso de uso do fluxo de trabalho

3. Configure o agente supervisor:
   - Nome: Nome descritivo do fluxo de trabalho
   - Descrição: Propósito e capacidades
   - Prompt do sistema: instruções para delegação e coordenação de agentes
   - Modelo: Modelo básico para o agente supervisor

Práticas recomendadas para solicitações do supervisor:

- Descreva claramente quando usar cada agente especializado
- Inclua instruções para agregar resultados de vários agentes
- Defina as expectativas de formatação da resposta
- Estabeleça limites para o comportamento da delegação

Seleção de agentes

Selecione agentes do Agent Builder para incluir como agentes especializados:

1. Clique em Adicionar agente na configuração do fluxo de trabalho
2. Procure ou pesquise agentes disponíveis do Agent Builder
3. Analise as descrições dos agentes
4. Selecione agentes para incluir no fluxo de trabalho

Descrições do agente

O agente supervisor usa as descrições dos agentes para decidir a qual agente delegar. Certifique-se de que as descrições expliquem claramente:

- Domínio ou capacidade especializada do agente
- Tipos de tarefas que o agente executa
- Expectativas de entrada/saída

Testando fluxos de trabalho

Após a implantação:

1. Acesse o fluxo de trabalho por meio do Painel de Implantação
2. Teste com consultas que exigem vários agentes
3. Monitore a delegação de agentes em CloudWatch registros
4. Analise a qualidade da resposta e os padrões de delegação
5. Ajuste o prompt do supervisor se a delegação estiver abaixo do ideal

Dicas para gerenciar os limites do token do modelo

Nota: A solução não tenta gerenciar diretamente os limites de tokens impostos por vários LLMs. Teste e garanta que sua solicitação permaneça dentro dos limites disponíveis impostos pelo fornecedor do modelo.

Para ajudar a controlar o tamanho dos prompts, tente o seguinte:

1. Familiarize-se com os limites impostos pelo modelo que você deseja usar. Esses valores podem diferir drasticamente entre os modelos, por isso é importante saber qual é o orçamento disponível antes de começar.
2. Crie sua solicitação inicial com esse orçamento em mente e considere quanto você deseja economizar em qualquer elemento dinâmico da solicitação. Por exemplo, entrada do usuário, histórico de bate-papo, trechos de documentos e assim por diante.
3. Na página de configuração do prompt, defina um limite para o tamanho do histórico final para limitar o número de turnos de conversação incluídos no prompt.
4. Defina limites de devolução de documentos no assistente de configuração da Base de Conhecimento. Você precisa tentar encontrar o equilíbrio certo entre fornecer ao LLM contexto suficiente para realizar a tarefa, mas não tanto a ponto de exceder os limites de token ou afetar negativamente a latência.
5. Deixe um pouco de amortecedor. Não faça um orçamento para um caso típico, pense e experimente casos extremos, como longas consultas de entrada, trechos grandes de documentos ou longas conversas.

Etapas para criar a imagem Docker do servidor MCP

Para usar servidores MCP (Model Context Protocol) com o Generative AI Application Builder na AWS, você precisa de uma imagem Docker criada e armazenada em um repositório privado do Amazon ECR como primeira etapa.

> **Note**
>
> No momento, os servidores MCP implantados existentes no tempo de AgentCore execução do Amazon Bedrock não podem ser exportados para o GAAB. Para que os servidores MCP sejam conectados aos agentes criados por meio do GAAB, eles precisam ser criados por meio do GAAB.

Etapa 1: Crie seu servidor MCP

Primeiro, você precisa ter a implementação do servidor MCP pronta. Para obter instruções detalhadas sobre como criar um servidor MCP, consulte o Amazon Bedrock AgentCore Developer Guide - Criar um servidor MCP.

Recomendamos a seguinte estrutura de projeto:

```
.
### __init__.py
### extras/
#   ### extra_dependencies.py
#   ### Dockerfile
### requirements.txt
### server.py <-- Server Entry point
```

Para a estrutura do Dockerfile, recomendamos usar um formato semelhante ao exemplo a seguir:

```
FROM ghcr.io/astral-sh/uv:python3.13-bookworm-slim
WORKDIR /app

# All environment variables in one layer
ENV UV_SYSTEM_PYTHON=1 \
    UV_COMPILE_BYTECODE=1 \
    UV_NO_PROGRESS=1 \
    PYTHONUNBUFFERED=1 \
    DOCKER_CONTAINER=1 \
    AWS_REGION=us-east-1 \
    AWS_DEFAULT_REGION=us-east-1

COPY requirements.txt requirements.txt
# Install from requirements file
```

```
RUN uv pip install -r requirements.txt

RUN uv pip install aws-opentelemetry-distro>=0.10.1

# Signal that this is running in Docker for host binding logic
ENV DOCKER_CONTAINER=1

# Create non-root user
RUN useradd -m -u 1000 bedrock_agentcore
USER bedrock_agentcore

EXPOSE 9000
EXPOSE 8000
EXPOSE 8080

# Copy entire project (respecting .dockerignore)
COPY . .

# Use the full module path
CMD ["opentelemetry-instrument", "python", "-m", "server"]
```

Etapa 2: Teste seu servidor MCP localmente

Antes de implantar na AWS, é importante testar seu servidor MCP localmente para garantir que ele funcione conforme o esperado. Para obter instruções detalhadas sobre testes locais, consulte o Amazon Bedrock AgentCore Developer Guide - Teste seu servidor MCP localmente.

Etapa 3: Implantar no Amazon ECR

Depois que seu servidor MCP for criado e testado localmente, siga estas etapas para implantá-lo no Amazon ECR:

1. Verifique se você tem a versão mais recente do AWS CLI e do Docker instalados. Para obter mais informações, consulte Conceitos básicos do Amazon ECR.
2. Recupere um token de autenticação e autentique seu cliente Docker em seu registro. Use a AWS CLI:

```
aws ecr get-login-password --region us-east-1 | docker login --username AWS --
password-stdin <account-id>.dkr.ecr.us-east-1.amazonaws.com
```

3. Crie sua imagem do Docker usando o comando a seguir. Para obter informações sobre como criar um arquivo Docker do zero, consulte a documentação do Docker. Você pode pular essa etapa se sua imagem já estiver criada:

```
docker build -t <repository-name> .
```

4. Depois que a compilação for concluída, marque sua imagem para que você possa enviá-la para esse repositório:

```
docker tag <repository-name>:latest <account-id>.dkr.ecr.us-east-1.amazonaws.com/
<repository-name>:latest
```

5. Execute o comando a seguir para enviar essa imagem para seu repositório AWS recém-criado:

```
docker push <account-id>.dkr.ecr.us-east-1.amazonaws.com/<repository-name>:latest
```

Para obter instruções completas de implantação, consulte o Amazon Bedrock AgentCore Developer Guide - Implante seu servidor MCP na AWS.

Etapa 4: usar o URI do ECR no GAAB

Depois de enviar com sucesso sua imagem do Docker para o Amazon ECR, copie o URI da imagem do console do ECR. Você usará esse URI ao implantar seu servidor MCP por meio do assistente de implantação Generative AI Application Builder na AWS.

Etapas para criar diferentes destinos do MCP Gateway

O Amazon Bedrock AgentCore Gateway permite que você transforme os serviços existentes da AWS APIs em ferramentas de MCP que podem ser usadas por seus agentes. O Gateway oferece suporte a vários tipos de destino, permitindo que você integre vários serviços de back-end sem problemas.

Os seguintes tipos de alvo são compatíveis:

- Metas do Lambda: transforme as funções do AWS Lambda em ferramentas MCP. Para obter instruções detalhadas, consulte o Guia do AgentCore desenvolvedor do Amazon Bedrock - Adicionar alvos Lambda.

- Destinos da OpenAPI: use as especificações da OpenAPI para definir e expor o REST como ferramentas MCP. APIs Para obter instruções detalhadas, consulte o Amazon Bedrock AgentCore Developer Guide - OpenAPI schema.
- Metas do Smithy: Crie ferramentas de MCP usando as definições do modelo Smithy para integrações de API com tipos seguros. Para obter instruções detalhadas, consulte o Amazon Bedrock AgentCore Developer Guide - Building Smithy targets.
- Destinos do MCP Server: conecte-se diretamente aos servidores MCP externos por meio de endpoints de URL, permitindo que você integre os servidores MCP existentes. Para obter instruções detalhadas, consulte o Amazon Bedrock AgentCore Developer Guide - destinos de servidores MCP.

Para ver exemplos e tutoriais adicionais sobre a criação de destinos do MCP Gateway, visite o repositório de amostras do Amazon Bedrock AgentCore .

Configurando uma base de conhecimento

Esta seção descreve como ingerir dados na base de conhecimento que você selecionou para a solução. Atualmente, a solução oferece suporte às bases de conhecimento Amazon Kendra e Amazon Bedrock como bases de conhecimento para sua implantação de casos de uso baseados em RAG.

Amazon Kendra

Se você estiver usando o Amazon Kendra como sua base de conhecimento, consulte o Guia do desenvolvedor do Amazon Kendra para obter informações sobre como usar vários conectores de fontes de dados para ajudá-lo a ingerir dados de uma ampla seleção de fontes.

Importante: para evitar a perda acidental de dados, a solução não exclui automaticamente o índice Kendra (criado pela solução ou não) quando uma implantação ou pilha é excluída. Se você quiser excluir sua base de conhecimento e parar de incorrer em custos, consulte a seção Desinstalação manual para obter detalhes sobre quais recursos são retidos e como limpá-los.

Bases de conhecimento do Amazon Bedrock

As bases de conhecimento Amazon Bedrock podem ser apoiadas por uma variedade de diferentes armazenamentos vetoriais, cada um com a capacidade de indexar seus dados. Para configurar e preencher sua base de conhecimento, consulte o Guia do usuário do Amazon Bedrock. Especificamente, você desejará:

- Primeiro, configure sua fonte de dados
- Em seguida, configure um índice vetorial para sua base de conhecimento em um repositório de vetores compatível. Observe que isso pode ser ignorado se você usar a opção “Criar rapidamente um novo repositório de vetores” no console Bedrock durante a criação da base de conhecimento.
- Por fim, você pode criar a base de conhecimento e sincronizar suas fontes de dados configuradas.

Configurações avançadas da base de conhecimento

Configurações avançadas da base de conhecimento, como filtragem da base de conhecimento e RAG com controle de acesso baseado em função, estão disponíveis para uso com a solução. A filtragem da base de conhecimento pode ser aplicada a qualquer uma das bases de conhecimento, enquanto o RAG com controle de acesso baseado em funções está disponível especificamente para o Amazon Kendra.

Filtragem da base de conhecimento

A solução permite que você especifique filtros de atributos do Amazon Kendra ou filtros de recuperação da base de conhecimento Bedrock ao implantar um caso de uso na seção Configurações avançadas do RAG da etapa da base de conhecimento do assistente. Esses filtros definem como as fontes de dados na base de conhecimento são consultadas, como estratégias de pesquisa, idiomas do documento subjacente que estão sendo consultados, etc.

Em ambos os casos, um objeto JSON é usado para especificar as configurações de filtro de acordo com o formato especificado na documentação de cada serviço (conforme vinculado acima).

Exemplo 1: Kendra AttributeFilter

```
{
 "EqualsTo": {
 "Key": "_language_code",
 "Value": {
 "StringValue": "es"
 }
 }
}
```

Exemplo 2: Bedrock RetrievalFilter

```
{
```

```
 "equals": {
 "key": "language",
 "value": "es"
 }
}
```

RAG com controle de acesso baseado em funções com Amazon Kendra

O controle de acesso baseado em funções (RBAC) permite controlar quais usuários ou grupos podem acessar determinados documentos em seu índice Amazon Kendra ou ver determinados documentos em seus resultados de pesquisa. Para configurar o RBAC para seu Amazon Kendra Index ID com seu caso de uso do Generative AI Application Builder na AWS (GAAB), siga estas etapas:

1. Configurar o Amazon Kendra Index

1. Certifique-se de ter um índice do Amazon Kendra criado e pelo menos uma fonte de dados adicionada a ele.
2. Configure o controle de acesso para sua fonte de dados com base nos grupos de usuários. Para uma fonte de dados do S3, siga as instruções disponíveis na documentação para configurar listas de controle de acesso (ACLs) usando os mesmos nomes de grupos criados no seu grupo de usuários do Amazon Cognito. Isso garante que os usuários só possam acessar os documentos e os resultados da pesquisa que estão autorizados a visualizar com base na associação ao grupo.

> **Note**
>
> Em Controle de acesso do usuário no Índice Kendra que você criou, deixe o controle de acesso do usuário baseado em token como Não. Quando você ativa o Controle de Acesso Baseado em Funções na Etapa 2, o Generative AI Application Builder na AWS extrai as declarações apropriadas do token de autenticação do usuário e cria um filtro de atributos.

2. Implante o caso de uso do RAG usando o assistente de implantação do GAAB

1. Siga as instruções do assistente na tela no Assistente de implantação do GAAB até chegar à etapa 4 do assistente para configurar o RAG.
2. Na etapa Selecionar base de conhecimento do assistente de implantação, escolha Amazon Kendra como o tipo de base de conhecimento.

3. Especifique se você tem um índice Amazon Kendra existente ou se deseja criar um novo. Se você tiver um índice existente, forneça o ID do seu índice Amazon Kendra que foi configurado com listas de controle de acesso ACLs () com base em grupos de usuários.
4. Ative a opção Controle de acesso baseado em função. Essa opção garante que os resultados da pesquisa retornados do índice Amazon Kendra sejam filtrados com base na função do usuário e nas permissões do grupo.
5. Analise e implante o caso de uso.

3. Configurar o Amazon Cognito

1. Localize o grupo de usuários do Amazon Cognito usado pela sua implantação do GAAB. Esse grupo de usuários do Amazon Cognito geralmente é criado pela pilha principal do painel CloudFormation de implantação.
2. Crie novos usuários no grupo de usuários do Amazon Cognito. Ao criar usuários, selecione a opção “Enviar um convite por e-mail” para que os usuários recebam credenciais de login temporárias por e-mail. Isso permite que novos usuários se inscrevam e acessem o aplicativo GAAB.
3. Crie grupos de usuários no grupo de usuários do Amazon Cognito. Certifique-se de que os nomes dos grupos correspondam exatamente aos grupos configurados no seu índice do Amazon ACLs Kendra. Isso é crucial para ativar o RBAC, pois a associação ao grupo do usuário determinará os resultados da pesquisa que ele poderá acessar.
4. Atribua usuários aos grupos apropriados com base em suas funções e permissões de acesso. Os usuários devem ser adicionados ao grupo necessário para a ACL do índice Amazon Kendra, bem como ao grupo específico do caso de uso criado durante a implantação do GAAB. Isso garante que os usuários tenham as permissões necessárias para acessar o caso de uso específico e os resultados de pesquisa relevantes.

Seguindo essas etapas, você terá configurado o controle de acesso baseado em função (RBAC) para sua implantação do GAAB, garantindo que os usuários só possam acessar e interagir com as informações e os recursos para os quais estão autorizados, com base no grupo de usuários e nas permissões atribuídas.

> **Note**
>
> No momento, somente o Amazon Kendra oferece suporte ao RBAC para bases de conhecimento no Generative AI Application Builder na AWS. Para o Amazon Bedrock

Knowledge Base, o RBAC não é suportado, mas você pode usar filtros de metadados para atingir algum nível de filtragem. Para obter mais informações, consulte o Guia do usuário do Amazon Bedrock.

Configurando seus prompts

O assistente do painel de implantação tem uma etapa de configuração imediata que permite que você personalize a experiência imediata e o modelo que orientará as interações entre os usuários e o modelo de IA. Definir adequadamente essas configurações é crucial para obter respostas precisas e relevantes do assistente de IA.

Esta seção controla a experiência geral e o comportamento do prompt de IA.

- Tamanho máximo do modelo de solicitação: essa configuração determina o tamanho máximo (em caracteres) do modelo de solicitação. Um valor mais alto permite que mais contexto seja fornecido ao modelo de IA, potencialmente levando a respostas mais precisas. No entanto, avisos excessivamente longos também podem introduzir ruído e afetar negativamente o desempenho. Para modelos Amazon Bedrock, os valores padrão para o tamanho máximo do modelo de solicitação (em caracteres) são calculados usando os limites do token do modelo subjacente. Se você editar e alterar o nome de um modelo no Bedrock, o botão “Redefinir para o padrão” será destacado e poderá ser usado para adotar os padrões do modelo recém-selecionado. Para modelos de SageMaker IA da Amazon, valores padrão razoáveis são fornecidos, mas é recomendável que você verifique seu modelo subjacente e escolha o tamanho máximo do modelo de solicitação e os comprimentos de texto de entrada de acordo. Consulte a seção Dicas para gerenciar os limites de tokens do modelo para obter mais informações.
- Tamanho máximo do texto de entrada: essa configuração limita o tamanho máximo (em caracteres) do texto de entrada do usuário. Entradas mais longas podem conter informações irrelevantes, aumentando o risco de obter respostas irrelevantes ou imprecisas do modelo de IA.
- Edição do prompt do usuário: essa opção permite que você ative ou desative a capacidade de os usuários modificarem o modelo do prompt por meio da interface do usuário do Chat. Desativar esse recurso pode ajudar a manter a consistência e evitar alterações não intencionais no prompt.

Modelo de prompt

Esta seção permite que você defina o modelo de prompt real que será usado pelo modelo de IA. O modelo de prompt normalmente segue uma estrutura que inclui espaços reservados para vários componentes, como a entrada do usuário, passagens de referência e histórico de bate-papo.

- Modelo de prompt: essa é a área de texto principal na qual você pode escrever ou colar o modelo de prompt desejado. O modelo deve ser criado para fornecer o contexto e as instruções necessários para o modelo de IA. Normalmente, inclui os seguintes espaços reservados:
  - `{input}`: esse espaço reservado é obrigatório para implantações do Sagemaker AI e será substituído pela entrada ou consulta do usuário.
  - `{history}`: Esse espaço reservado é obrigatório para implantações do Sagemaker AI e será substituído pelo histórico de bate-papo da conversa atual.
  - `{context}`: Esse espaço reservado é obrigatório para implantações do RAG e será substituído pelos trechos do documento obtidos da base de conhecimento configurada.
- Reformular a pergunta? : essa opção (disponível somente para implantações de RAG) determina se a consulta de entrada original do usuário deve ser reformulada ou desambiguada antes de ser passada para o modelo de IA. Às vezes, reformular a consulta pode ajudar o modelo a entender melhor a intenção do usuário, o que pode levar a respostas mais precisas.

Ao configurar o modelo e a experiência do prompt, é essencial encontrar um equilíbrio entre fornecer contexto e instruções suficientes ao modelo de IA e, ao mesmo tempo, evitar informações excessivamente longas ou irrelevantes que possam causar problemas de ruído ou desempenho.

Configurações avançadas de prompt

Esta seção permite controlar como o histórico de conversas é apresentado ao modelo de IA.

- Tamanho do histórico final: essa configuração determina o número de mensagens anteriores que devem ser incluídas no prompt final. Definir esse valor como zero resultaria em nenhum histórico sendo injetado no modelo de prompt ou no modelo de prompt de desambiguação. Observação: mesmo quando definido como zero, ainda é necessário que exista um espaço reservado {history} nos modelos de prompt. Em tempo de execução, ele será substituído por uma string vazia.
  - Nota: É recomendável fornecer um número par para esse valor. Fornecer um número ímpar resultaria no retorno apenas da resposta de IA de uma interação pareada.
- Prefixo humano: é o prefixo usado para identificar as mensagens enviadas pelo usuário no histórico de conversas.

- Prefixo de IA: é o prefixo usado para identificar mensagens retornadas pelo modelo de IA no histórico de conversas.

Configuração do prompt de desambiguação

Esta seção permite configurar o comportamento e o modelo para eliminar a ambiguidade das entradas do usuário antes de enviá-las para a base de conhecimento configurada.

- Ativar desambiguação: essa opção determina se as entradas do usuário devem ser desambiguadas antes de serem enviadas para a base de conhecimento.
- Modelo de aviso de desambiguação: Esse é o modelo de aviso usado para eliminar a ambiguação das entradas do usuário quando conectado a uma base de conhecimento. A saída gerada a partir desse prompt será usada como a consulta enviada à base de conhecimento. A desativação da desambiguação resultaria no envio inalterado da consulta bruta do usuário para a base de conhecimento.

Por exemplo, com a desambiguação ativada, um usuário de acompanhamento pergunta “Quanto custa?” pode ser desambiguado para “Quanto custa renovar minha placa?” , levando a uma melhor consulta de pesquisa.

Usando o caso de uso do Text implantado

A interface de usuário integrada para o caso de uso do Text tem como objetivo permitir que os usuários corporativos explorem e experimentem rapidamente a implantação criada pelo usuário administrador. As alterações de configuração feitas pelo usuário corporativo só entram em vigor em sua sessão. O usuário corporativo deve compartilhar essas alterações com o usuário administrador, que pode atualizar a implantação básica com essas alterações para que todos possam usar.

A interface do usuário do chat consiste nos seguintes componentes:

- Janela de bate-papo
- Caixa de entrada de bate-papo
- Configurações
- Conversa clara

Janela de bate-papo

Tem diferentes turnos da conversa. As mensagens que começam à direita são do usuário corporativo e as mensagens que começam à esquerda são do LLM configurado. Existe um pequeno ícone de prancheta em todas as respostas do LLM para facilitar a cópia das respostas.

Caixa de entrada de bate-papo

Fixada na parte inferior da janela de bate-papo está a caixa de entrada do bate-papo. É aqui que os usuários corporativos podem inserir suas mensagens para serem enviadas ao LLM. Logo acima da caixa de entrada está o status da conexão. Se a conexão for perdida (por exemplo, devido à inatividade), uma nova conexão será criada automaticamente na próxima vez que uma mensagem de bate-papo for enviada. Espera-se que essa solicitação demore um pouco mais devido ao tempo adicional de WebSocket conexão.

Com base na configuração específica, pode haver um comprimento máximo imposto na entrada. Se esse limite for excedido, os usuários receberão um alerta e a mensagem não será enviada.

Observação: se estiver usando o RAG com o Amazon Kendra, a API Retrieve truncará as consultas para 30 palavras simbólicas. Se estiver esperando mais entradas do usuário, avalie como isso pode afetar o desempenho da pesquisa.

Configurações

Para permitir que os usuários corporativos experimentem rapidamente diferentes configurações, um painel de configurações está disponível, o que permite a on-the-fly edição de determinadas opções de configuração de implantação

(exemplo, modelo de prompt). Essas alterações só podem ser feitas no início de uma nova sessão. Depois que uma conversa é iniciada, a limpeza da conversa reativa a edição das configurações.

Observação: os usuários administradores podem optar por bloquear as configurações de uma implantação. Eles podem evitar edições ao vivo no momento da implantação por meio do assistente durante a etapa de solicitação.

Conversa clara

Ao longo da conversa, a solução mantém um histórico de bate-papo, o que permite uma experiência de conversação. Isso permite a desambiguação da consulta e perguntas complementares. Para

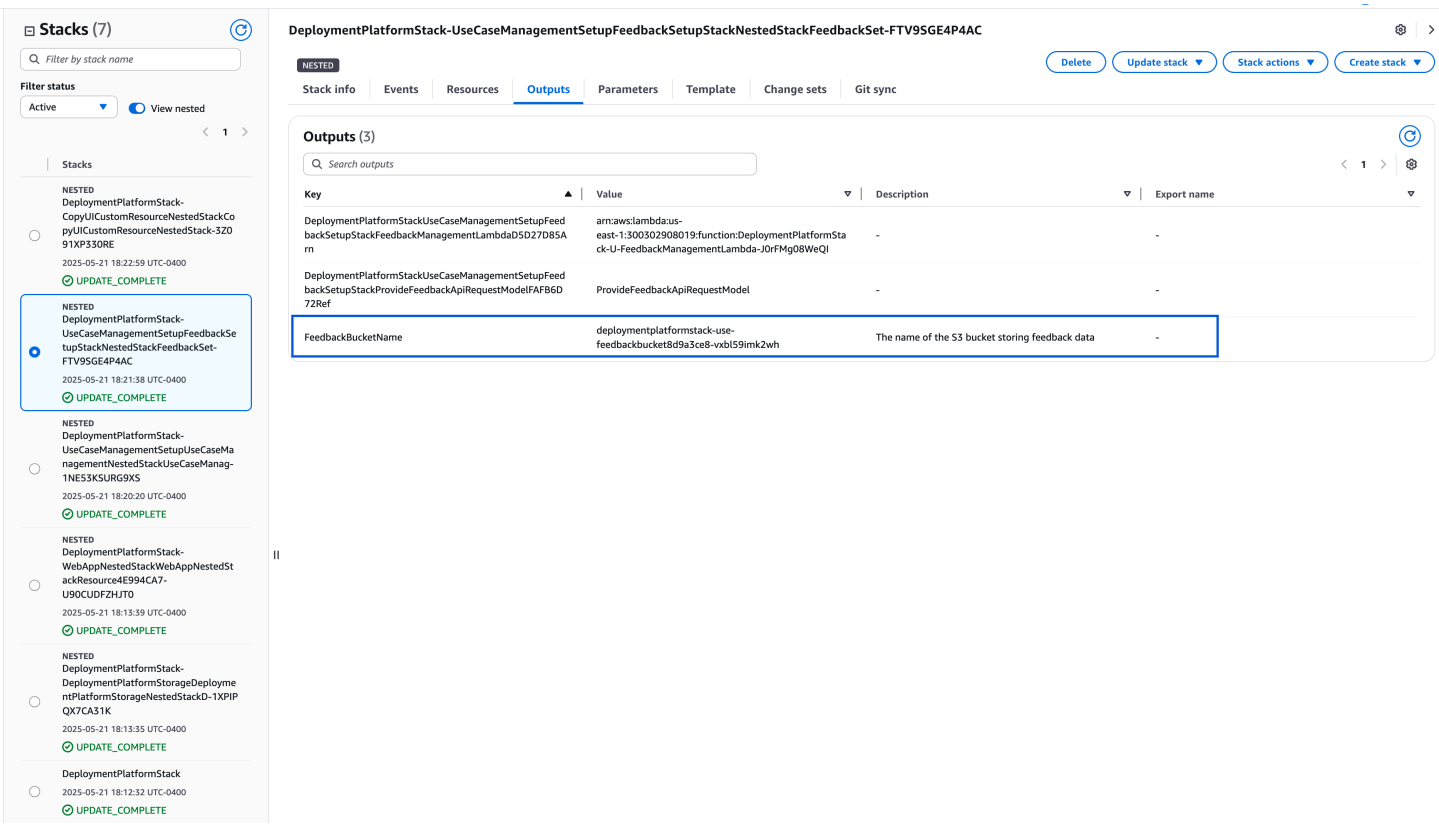
redefinir uma conversa e excluir todo o histórico de bate-papo dessa interação, escolha *Limpar conversa* na parte superior da janela de bate-papo. Depois que a conversa for encerrada, uma nova sessão será criada, reativando a edição das configurações.

Acessando e analisando o feedback coletado pelo usuário

A partir da versão 3.0.0, o Painel de Implantação implanta uma pilha de feedback aninhada que permite que os casos de uso do Text e do Bedrock Agent implantados com o Painel tenham a funcionalidade de coleta de feedback para as respostas geradas. LLM/Agent Particularmente, os usuários podem fornecer um feedback positivo ou negativo junto com um comentário opcional. Se o usuário fornecer um feedback negativo, ele poderá selecionar ainda mais uma dessas categorias negativas: 'Impreciso', 'Incompleto ou insuficiente', 'Nocivo' 'Outro'. and/or

Depois que o usuário fornece o feedback, o feedback é armazenado em um bucket do S3 particionado por ID de caso de uso, ano e mês. O ID do caso de uso pode ser encontrado no Painel de implantação e o bucket do Feedback S3 pode ser encontrado nas saídas da pilha aninhada de feedback da pilha do Painel de implantação:

Descreve a pilha de implantação - Finding Feedback Bucket Name



O feedback do usuário é enviado como uma solicitação de API contendo um conjunto mínimo de informações:

```
{
  "useCaseRecordKey": "a1b2c3d4-e5f6g7h8",
  "conversationId": "12345678-1234-1234-1234-123456789012",
  "messageId": "87654321-4321-4321-4321-210987654321",
  "rephrasedQuery": "What are the key features of the Generative AI Application Builder
 on AWS?",
  "sourceDocuments": [
    "s3://bucket-name/document1.pdf",
    "s3://bucket-name/document2.pdf"
  ],
  "feedback": "positive",
  "feedbackReason": [
    "Incomplete or insufficient"
  ],
  "comment": "The response was helpful but could include more details about important
 features."
}
```

Essa carga é então processada por um lambda usando o `useCaseRecordKey` que identifica a configuração correta de um caso de uso no momento da implantação. Essa configuração é usada para obter detalhes específicos do feedback, como o nome ConversationTable do (contém todas as conversas e sequências de mensagens humanas e de IA), que é usado posteriormente para recuperar o e. `userInput` `llmResponse` Detalhes adicionais também estão anexados a esse registro de feedback, como `agentId` e `agentAliasId` para um caso de uso do Bedrock Agent e `modelProviderbedrockModelId`, etc. para um caso de uso de texto usando essa configuração. Para obter detalhes sobre como acessar essa configuração, consulte a seção Mapeamentos de feedback personalizados abaixo. Cada solicitação de feedback recebida é armazenada como um objeto JSON e um exemplo de registro de feedback pode ter a seguinte aparência para um caso de uso de texto:

```
{
   "useCaseId": "12345678-1234-1234-1234-123456789012",
   "useCaseRecordKey": "c07a2e3b-2f31b1e0",
   "userId": "22345678-1234-1234-1234-123456789012",
   "conversationId": "dd51de5d-5af1-4ec6-91d2-aadf14352109",
   "messageId": "32345678-1234-1234-1234-123456789012",
   "userInput": "What are its key features?",
```

```
   "rephrasedQuery": "What are the key features of the Generative AI Application
 Builder on AWS?",
   "llmResponse": "Generative AI Application Builder on AWS can help you build
 production ready enterprise chatbots rapidly.",
   "feedback": "negative",
   "feedbackReason": [
      "Incomplete or insufficient"
   ],
   "comment": "The response was helpful but could include more details about important
 features.",
   "timestamp": "2025-05-22T18:48:08.340Z",
   "feedbackId": "42345678-1234-1234-1234-123456789012",
   "useCaseType": "Text",
   "modelProvider": "Bedrock",
   "bedrockModelId": "amazon.nova-lite-v1:0",
   "ragEnabled": "false"
}
```

ou assim para um caso de uso do Bedrock Agent:

```
{
   "useCaseId": "12345678-1234-1234-1234-123456789012",
   "useCaseRecordKey": "c07a2e3b-2f31b1e0",
   "userId": "22345678-1234-1234-1234-123456789012",
   "conversationId": "dd51de5d-5af1-4ec6-91d2-aadf14352109",
   "messageId": "32345678-1234-1234-1234-123456789012",
   "userInput": "What are its key features?",
   "llmResponse": "Generative AI Application Builder on AWS can help you build
 production ready enterprise chatbots rapidly.",
   "feedback": "negative",
   "feedbackReason": [
      "Incomplete or insufficient"
   ],
   "comment": "The response was helpful but could include more details about important
 features.",
   "timestamp": "2025-05-22T18:48:08.340Z",
   "feedbackId": "42345678-1234-1234-1234-123456789012",
   "useCaseType": "Agent",
   "agentId": "AHFXUJCAK1",
   "agentAliasId": "KSEDKOS0BL"
}
```

Esse feedback pode então ser usado para processamento adicional, análise e modelagem de ciclos de retreinamento/feedback. Você também pode adicionar mapeamentos personalizados para aprimorar o registro de feedback que está sendo armazenado no feedback lambda.

Mapeamentos de feedback personalizados

O Painel de Implantação contém um `LLMConfigTable` que pode ser encontrado nas saídas da pilha do Painel de Implantação com a chave. `LLMConfigTableName` `LLMConfigTable`contém as configurações para cada caso de uso com base nas configurações selecionadas pelo administrador ao implantar o caso de uso por meio do assistente do Deployment Dashboard. Cada configuração de caso de uso é identificada por sua. `useCaseRecordKey` Aqui está um exemplo de registro de configuração de caso de uso no: `LLMConfigTable`

```
{
   "key": "2dd76cfa-bc1a14da",
   "config": {
      "ConversationMemoryParams": {
        ...
      },
      "FeedbackParams": {
         "CustomMappings": {
            "NumberOfDocs": "$.KnowledgeBaseParams.NumberOfDocs",
            "ScoreThreshold": "$.KnowledgeBaseParams.ScoreThreshold"
         },
         "FeedbackEnabled": true
      },
      "IsInternalUser": "true",
      "KnowledgeBaseParams": {
         "KendraKnowledgeBaseParams": {
            "ExistingKendraIndexId": "d2831033-667f-4539-ab28-e6c7c7c5988b",
            "RoleBasedAccessControlEnabled": false
         },
         "KnowledgeBaseType": "Kendra",
         "NumberOfDocs": 5,
         "ReturnSourceDocs": false,
         "ScoreThreshold": 0.3
      },
      "LlmParams": {
         "BedrockLlmParams": {
            "BedrockInferenceType": "QUICK_START",
            "ModelId": "amazon.nova-lite-v1:0"
         },
```

```
            "ModelParams": {},
            "ModelProvider": "Bedrock",
            "PromptParams": {
               ...
            },
            "RAGEnabled": true,
            "Streaming": false,
            "Temperature": 0.1,
            "Verbose": false
         },
         "UseCaseName": "test-rag-usecase",
         "UseCaseType": "Text"
      }
}
```

Se o feedback estiver habilitado para um caso de uso, essa configuração conterá um `FeedbackParams` objeto que permite que um `CustomMappings` objeto dentro dela possa especificar que todos JSONPaths os campos adicionais sejam adicionados ao registro JSON de feedback armazenado no bucket de feedback do S3. Por exemplo, para o exemplo de configuração de caso de uso acima, o CustomMappings contém `NumberOfDocs` e, `ScoreThreshold` JSONPaths adicionalmente, no `CustomMappings` objeto que começa com `config` como raiz do. JSONPath Com essa configuração, cada registro JSON armazenado no bucket de feedback do S3 começará a receber esses dois valores adicionais além dos campos que já foram fornecidos.

Analisando dados de feedback

Os dados de feedback são armazenados no S3 como objetos JSON. Aqui estão algumas abordagens para tornar esses dados de feedback mais acessíveis e acionáveis:

Usando o AWS Glue e o Amazon Athena

O AWS Glue e o Amazon Athena oferecem uma forma sem servidor de catalogar, consultar e analisar seus dados de feedback.

O AWS Glue permite que você crie um crawler do AWS Glue que inspeciona os dados em um bucket do S3, infere seu esquema e registra todos os metadados relevantes em um catálogo. Depois disso, serviços como o Amazon Athena podem ser usados para consultar os dados.

Você pode consultar a documentação do AWS Athena sobre as etapas para conectar o bucket S3 de feedback ao Amazon Athena usando o AWS Glue Data Catalog. Você também pode usar alguns dos

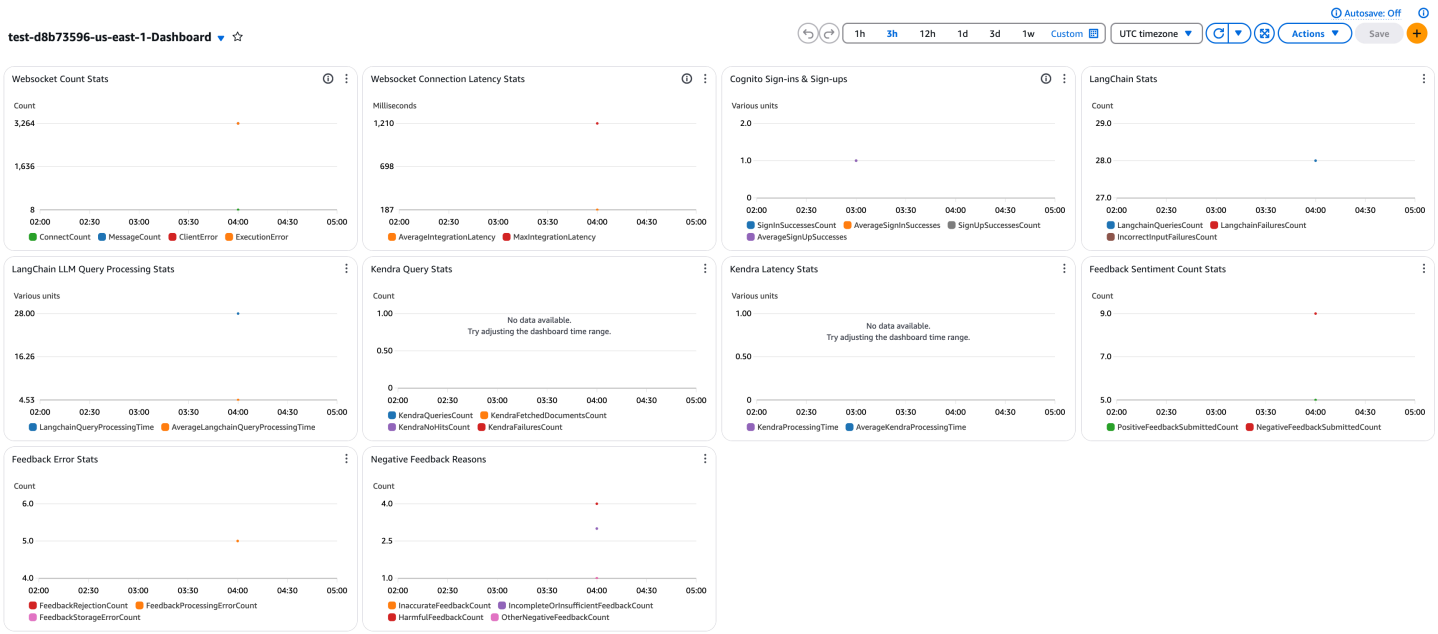
recursos mais poderosos do Glue para realizar trabalhos de extração, transformação e carregamento (ETL) nesses dados e transformá-los em um formato adequado aos seus casos de uso de análise ou retreinamento de modelos. Com o Glue, você pode realizar operações como filtrar os registros com determinados tipos de feedback, preencher qualquer informação ausente e também pode carregar esses dados em outro local de armazenamento, como outro bucket do S3 ou outro armazenamento de dados da AWS.

> **Note**
>
> Dependendo do seu caso de uso, considere programar o rastreador Glue para ser executado periodicamente (por exemplo, semanalmente) em vez de todas as noites para otimizar os custos, pois os dados de feedback podem ser escassos.

Usando os CloudWatch painéis da solução

Você também tem acesso a um CloudWatch painel com a solução que pode fornecer tendências de feedback positivo e negativo, categorias de motivos de feedback negativo, etc. de acordo com o caso de uso. Você pode encontrar esse painel usando o nome do seu caso de uso em Painéis dentro do console da AWS: CloudWatch

Descreve o painel de casos de uso CloudWatch



Você também pode criar widgets adicionais neste painel ou criar painéis do Amazon Quick Sight.

Melhores práticas para análise de dados de feedback

- Implemente políticas de ciclo de vida de dados em seu bucket S3 para arquivar dados de feedback mais antigos em níveis de armazenamento de menor custo
- Crie análises separadas para cada caso de uso para identificar oportunidades de melhoria específicas do modelo
- Estabeleça limites de feedback que acionem alertas quando o feedback negativo excede os níveis aceitáveis
- Exporte informações críticas periodicamente para compartilhar com as partes interessadas e as equipes de melhoria do modelo

Visualizando métricas de operação para uma implantação

O painel de implantação e as pilhas de casos de uso vêm com seu próprio CloudWatch painel, rastreando várias métricas operacionais da solução. Você pode usar esses CloudWatch painéis para ajudar a comparar diferentes implantações. Para acessar os painéis:

1. Navegue até o console do CloudWatch .
2. Pesquise os painéis pré-criados pesquisando o nome da pilha ou o identificador exclusivo universal (UUID).

Por exemplo, o caso de uso do Text vem com gráficos que rastreiam o número de WebSocket conexões, o número de logins e inscrições de usuários, a quantidade de tempo que o LLM levou para processar uma conclusão e assim por diante. Os clientes podem usar esses gráficos para comparar várias _métricas _quantitativas de uma implantação.

Example

É difícil comparar os resultados qualitativos de vários modelos aplicados a diferentes casos de uso. Use o recurso Clone para acelerar várias implantações rapidamente para que você possa comparar as saídas lado a lado.

Informações sobre CloudWatch registros de acesso

Essa solução registra mensagens de erro, aviso, informações e depuração para as funções do Lambda. Para escolher o tipo de mensagem a ser registrada:

1. Localize a função aplicável no console do AWS Lambda.
2. Adicione uma variável de ambiente POWERTOOLS_LOG_LEVEL.
3. Defina a variável para o tipo de mensagem aplicável.

Para obter mais instruções, consulte Criar variáveis de ambiente do Lambda no Guia do desenvolvedor do AWS Lambda.

A tabela a seguir lista os tipos de níveis de log que você pode escolher.

Nível	Description
ERROR (Erro)	Os registros incluem informações sobre qualquer coisa que faça com que uma operação falhe.
AVISO	Os registros incluem informações sobre qualquer coisa que possa causar inconsist ências na função, mas não necessariamente causar falha na operação. Os registros também incluem mensagens de ERRO.
INFO	Os registros incluem informações de alto nível sobre como a função está operando. Os registros também incluem mensagens de ERRO e AVISO.
DEBUG	Os registros incluem informações que podem ser úteis ao depurar um problema com a função. Os registros também incluem mensagens de ERRO, AVISO e INFORMAÇÕ ES.

Use o procedimento a seguir para adicionar insights do CloudWatch Logs a essa solução.

1. Identifique os grupos de registros relevantes:
   a. Faça login no CloudFormation console da AWS.
   b. Escolha sua pilha alvo.

c. Selecione a guia Recursos e pesquise suas funções Lambda de destino.

d. Faça login no console do AWS Lambda e escolha cada uma das suas funções de destino do Lambda.

e. Para cada uma das funções do Lambda de destino, selecione a guia Monitor e escolha Exibir CloudWatch registros.

f. Copie os nomes dos grupos de registros dos quais você deseja extrair insights.

2. Navegue até o CloudWatch console da Amazon.
3. No menu de navegação, em Registros, escolha Logs Insights.
4. Na página Logs Insights, escolha a guia Registros.
5. Pesquise nomes de grupos de registros na etapa 1.
6. Copie uma das consultas de exemplo a seguir e cole-a no campo de consulta:

a. Para identificar todas as exceções do cliente:

```
fields @message
|filter @message like /(?i)Exception/|stats count(*) as exceptionCount by @message
```

b. Para recuperar a contagem de invocações pelo nome da função:

```
stats count(*) by function_name
```

c. Para recuperar a contagem de invocações em intervalos de cinco minutos:

```
stats count(*) as invocations by bin(5m)
```

d. Para recuperar todo o rastreamento IDs do AWS X-Ray:

```
filter @message like "XRAY TraceId"
|parse @message "XRAY TraceId: * " as traceId|stats count(*) by traceId
```

e. Para recuperar registros relacionados a um X-Ray Trace ID específico:

```
filter @message like "your-traceid-here"
```

f. Para recuperar erros não autorizados WebSocket :

```
fields
@ingestionTime,
@log,
```

```
@logStream,
@message,
@requestId,
@timestamp,
errorMessage,
errorType
|filter @message like /Unauthorized/ and @message like /websocket/|sort @timestamp
 desc
```

g. Para recuperar a contagem de métricas publicadas:

```
filter @message like "CloudWatchMetrics"
|parse @message /"Metrics":\s*\[(?<metrics>.*?)\]/|stats count(*) as metric_count
 by metrics
```

Guia do desenvolvedor

Esta seção fornece o código-fonte da solução, um guia de integração, um guia de personalização e uma referência de API.

Código-fonte

Visite nosso GitHub repositório para baixar os arquivos de origem dessa solução e compartilhar suas personalizações com outras pessoas.

Os modelos do Generative AI Application Builder na AWS são gerados usando o AWS Cloud Development Kit (AWS CDK). Consulte o arquivo README.md para obter informações adicionais.

Guia de integração

Toda a solução foi projetada para ser facilmente extensível. A camada de orquestração dessa solução é criada usando. LangChain Você pode adicionar qualquer provedor de modelo, base de conhecimento ou tipo de memória de conversação suportado por LangChain (ou um terceiro que forneça LangChain conectores para esses componentes) a essa solução.

Suporte de expansão LLMs

Para adicionar outro provedor de modelo, como um provedor LLM personalizado, você deve atualizar os três componentes da solução a seguir:

1. Crie uma nova pilha de `TextUseCase` CDK, que implanta o aplicativo de bate-papo configurado com seu provedor LLM personalizado:
   a. Clone o GitHub repositório dessa solução e configure seu ambiente de compilação seguindo as instruções fornecidas no arquivo README.md.
   b. Copie (ou crie um novo) o `source/infrastructure/lib/bedrock-chat-stack.ts` arquivo, cole-o no mesmo diretório e renomeie-o para`custom-chat-stack.ts`.
   c. Renomeie a classe no arquivo para uma adequada, como`CustomLLMChat`.
   d. Você pode optar por adicionar um segredo do Secrets Manager a essa pilha, que armazena suas credenciais para seu LLM personalizado. Você pode recuperar essas credenciais durante a invocação do modelo na camada Lambda de bate-papo discutida no próximo parágrafo.

2. Crie e anexe uma camada Lambda contendo a biblioteca Python do provedor de modelo a ser adicionada. Para um aplicativo de bate-papo de casos de uso do Amazon Bedrock, a biblioteca `langchain-aws` Python contém os conectores personalizados na parte superior do pacote para se conectar aos provedores de LangChain modelos da AWS (Amazon Bedrock SageMaker e AI), às bases de conhecimento (Amazon Kendra e Amazon Bedrock Knowledge Bases) e aos tipos de memória (como o DynamoDB). Da mesma forma, outros fornecedores de modelos têm seus próprios conectores. Essa camada ajuda você a anexar a biblioteca Python desse provedor de modelos para que você possa usar esses conectores na camada Lambda de bate-papo, que invoca o LLM (etapa 3). Nessa solução, um agrupador de ativos personalizado é usado para criar camadas Lambda, que são anexadas usando aspectos do CDK. Para criar uma nova camada para a biblioteca do provedor de modelos personalizados:

   a. Navegue até a `LambdaAspects` classe no `source/infrastructure/lib/utils/lambda-aspects.ts` arquivo.

   b. Siga as instruções sobre como estender a funcionalidade da classe de aspectos Lambda fornecida no arquivo (como adicionar o `getOrCreateLangchainLayer` método). Para usar esse novo método (por exemplo,`getOrCreateCustomLLMLayer`), atualize também a `LLM_LIBRARY_LAYER_TYPES` enumeração no `source/infrastructure/lib/utils/constants.ts` arquivo.

3. Estenda a função `chat` Lambda para implementar um construtor, cliente e manipulador para o novo provedor.

   O `source/lambda/chat` contém as LangChain conexões de diferentes classes, LLMs juntamente com as classes de suporte para construí-las LLMs. Essas classes de suporte seguem os padrões de design Builder e Object Oriented para criar o LLM.

   Cada manipulador (por exemplo,`bedrock_handler.py`) primeiro cria um cliente, verifica o ambiente em busca das variáveis de ambiente necessárias e, em seguida, chama um `get_model` método para obter a classe LangChain LLM. O método generate é então chamado para invocar o LLM e obter sua resposta. LangChain atualmente oferece suporte à funcionalidade de streaming para o Amazon Bedrock, mas não para SageMaker IA. Com base na funcionalidade de streaming ou não streaming, o WebSocket manipulador apropriado (`WebsocketStreamingCallbackHandler`ou`WebsocketHandler`) é chamado para enviar a resposta de volta à WebSocket conexão usando o `post_to_connection` método.

   A `clients/builder` pasta contém as classes que ajudam a criar um LLM Builder usando o padrão Builder. Primeiro, a `use_case_config` é recuperado de um armazenamento de configurações do DynamoDB, que armazena os detalhes sobre o tipo de base de conhecimento,

memória de conversação e modelo a ser construído. Ele também contém detalhes relevantes do modelo, como parâmetros e avisos do modelo. Em seguida, o Builder ajuda a seguir as etapas para criar uma base de conhecimento, criar uma memória de conversação para manter o contexto da conversa para o LLM, definir os retornos de LangChain chamada apropriados para casos de streaming e não streaming e criar um modelo LLM com base nas configurações de modelo fornecidas. A configuração do DynamoDB é armazenada no momento da criação do caso de uso, quando você implanta um caso de uso a partir do painel de implantação (ou quando é fornecida pelos usuários em implantações autônomas de pilha de casos de uso sem o painel de implantação).

A `clients/factories` subpasta ajuda a definir a memória de conversação e a classe da base de conhecimento apropriadas, com base na configuração do LLM. Isso permite uma fácil extensão para qualquer outra base de conhecimento ou tipo de memória que você queira que sua implementação ofereça suporte.

A `shared` subpasta contém implementações específicas da base de conhecimento e da memória de conversação que são instanciadas dentro das fábricas pelo construtor. Ele também contém recuperadores do Amazon Kendra e do Amazon Bedrock Knowledge Base LangChain chamados para recuperar documentos para os casos de uso do RAG, além de retornos de chamada, que são usados pelo modelo LLM. LangChain

As LangChain implementações usam a Linguagem de LangChain Expressão (LCEL) para compor cadeias de conversação. `RunnableWithMessageHistory`A classe é usada para manter o histórico de conversas com cadeias de LCEL personalizadas, permitindo que funcionalidades como retornar documentos de origem e usar a pergunta reformulada (ou desambiguada) enviada à base de conhecimento também sejam enviadas ao LLM.

Para criar sua própria implementação de um provedor personalizado, você pode:

a. Copie o `bedrock_handler.py` arquivo e crie seu manipulador personalizado (por exemplo,`custom_handler.py`), que cria seu cliente personalizado (por exemplo,`CustomProviderClient`) (especificado na etapa a seguir).
b. Copie `bedrock_client.py` na pasta de clientes. Renomeie-o para `custom_provider_client.py` (ou para o nome específico do provedor do modelo, como`CustomProvider`). Nomeie a classe dentro dela de forma adequada, como a `CustomProviderClient` que `LLMChatClient` herda.

   Você pode usar os métodos fornecidos por `LLMChatClient` ou escrever suas próprias implementações para substituí-los.

O `get_model` método cria um `CustomProviderBuilder` (consulte a etapa a seguir) e chama o `construct_chat_model` método que constrói o modelo de bate-papo usando as etapas do construtor. Esse método atua como diretor no padrão do construtor.

c. Copie `clients/builders/bedrock_builder.py` e renomeie para `custom_provider_builder.py` e a classe dentro dela para `CustomProviderBuilder` que herda LLMBuilder ()`llm_builder.py`. Você pode usar os métodos fornecidos por LLMBuilder ou escrever suas próprias implementações para substituí-los. As etapas do construtor são chamadas em sequência dentro do `construct_chat_model` método do clienteset_model_defaults, comoset_knowledge_base, `set_conversation_memory` e.

   O `set_llm_model` método criaria o modelo LLM real usando todos os valores definidos usando os métodos chamados antes dele. Especificamente, você pode criar um LLM RAG (`CustomProviderRetrievalLLM`) ou não RAG (`CustomProviderLLM`), com base no `rag_enabled variable` que é recuperado da configuração do LLM no DynamoDB.

   Essa configuração é obtida no `retrieve_use_case_config` método da `LLMChatClient` classe.

d. Implemente sua `CustomProviderRetrievalLLM` implementação `CustomProviderLLM` ou na `llm_models` subpasta com base na necessidade de um caso de uso RAG ou não RAG. A maioria das funcionalidades para implementar esses modelos é fornecida em suas `RetrievalLLM` classes `BaseLangChainModel` e, respectivamente, para casos de uso não RAG e RAG.

   Você pode copiar o `llm_models/bedrock.py` arquivo e fazer as alterações necessárias para chamar o LangChain modelo que se refere ao seu provedor personalizado. Por exemplo, o Amazon Bedrock usa uma `ChatBedrock` classe para criar um modelo de bate-papo usando LangChain.

   O método generate gera a resposta LLM usando as cadeias LangChain LCEL.

   Você também pode usar o `get_clean_model_params` método para higienizar os parâmetros do modelo de acordo LangChain com os requisitos do seu modelo.

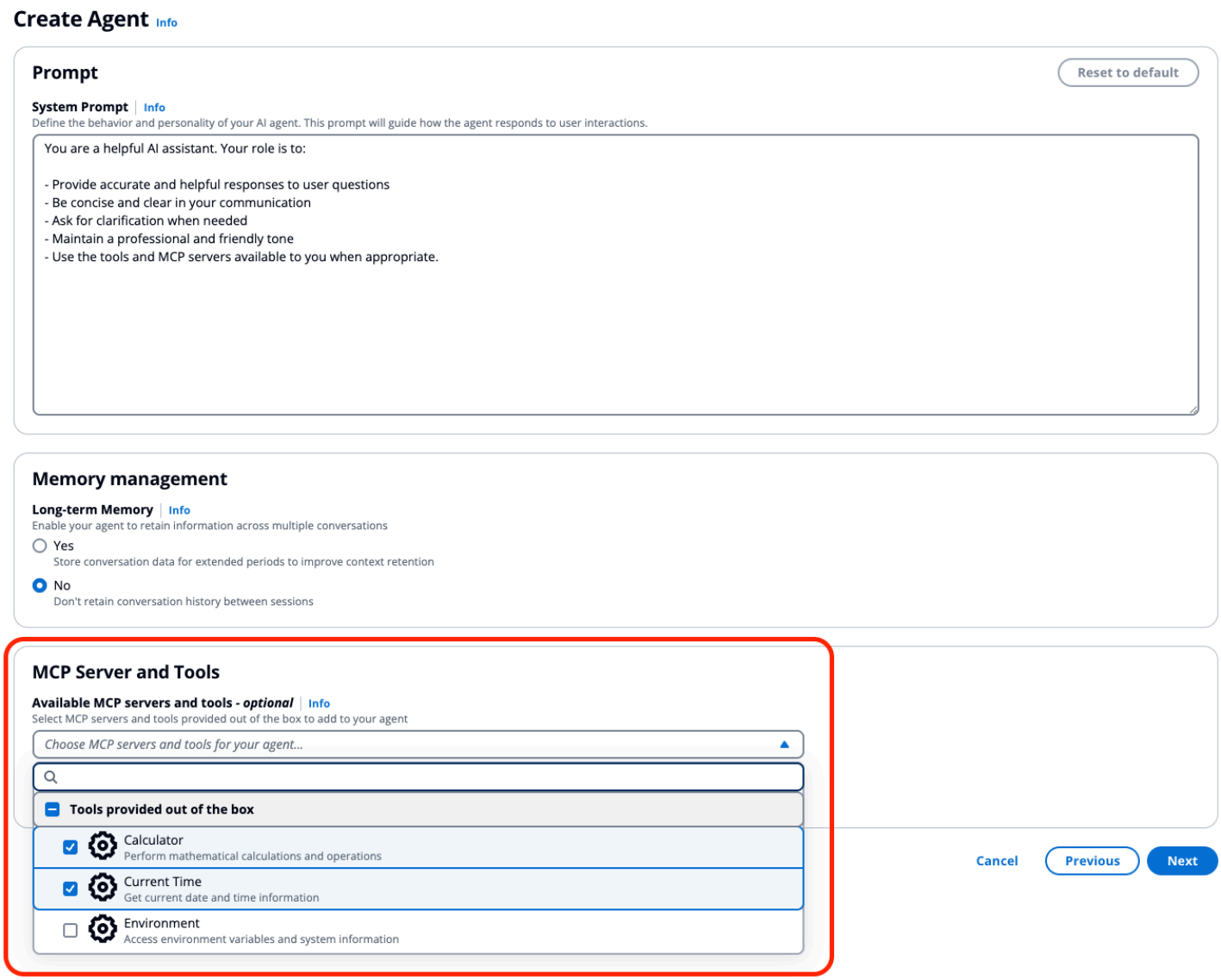
Expandindo as ferramentas Strands suportadas

A solução permite que você crie e implante servidores MCP, agentes de IA e fluxos de trabalho com vários agentes. Na experiência do Agent Builder, você pode conectar servidores MCP para oferecer recursos adicionais aos seus agentes. Além dos servidores MCP, você pode aproveitar as ferramentas integradas fornecidas pela Strands (a estrutura subjacente usada pela solução).

Pronta para uso, a solução vem pré-configurada com as seguintes ferramentas Strands:

- Hora atual (ativada por padrão)
- Calculadora (ativada por padrão)
- Environment

Seleção de servidor e ferramentas MCP no assistente do Agent Builder mostrando as ferramentas Strands integradas



Para ampliar seus agentes com ferramentas adicionais da Strands, siga o processo de quatro etapas descrito nesta seção.

Etapa 1: Encontre a ferramenta Strands

Navegue pelas ferramentas Strands disponíveis para identificar a ferramenta que você deseja usar. Cada ferramenta tem recursos e requisitos de configuração específicos.

Por exemplo, para adicionar recursos de recuperação da Base de Conhecimento Amazon Bedrock, você usaria a ferramenta de recuperação.

Etapa 2: atualizar o parâmetro SSM

Para disponibilizar uma ferramenta na interface de implantação do Agent Builder, atualize o parâmetro AWS Systems Manager Parameter Store que define quais ferramentas Strands são suportadas.

1. Navegue até o AWS Systems Manager Parameter Store em sua conta da AWS.
2. Localize o parâmetro: `/gaab/<stack-name>/strands-tools`
3. Adicione a configuração da ferramenta ao final da lista existente usando a seguinte estrutura JSON:

```
{
  "name": "Bedrock KB Retrieve",
  "description": "Retrieve information from Bedrock Knowledge Base",
  "value": "retrieve",
  "category": "AI",
  "isDefault": false
}
```

Campo	Description
name	Nome de exibição mostrado na interface do Agent Builder
descrição	Breve descrição da funcionalidade da ferramenta
value	O nome exato da ferramenta, conforme definido no pacote de ferramentas Strands
category	Categoria organizacional para agrupar ferramentas na interface do usuário
é o padrão	Se a ferramenta deve ser ativada por padrão para novos agentes

Etapa 3: Configurar variáveis de ambiente

Muitas ferramentas Strands exigem variáveis de ambiente para configuração. Você pode definir essas variáveis de duas maneiras:

Opção 1: configuração direta no AgentCore Runtime

Atualize o agente implantado diretamente no Amazon Bedrock AgentCore Runtime com as variáveis de ambiente necessárias.

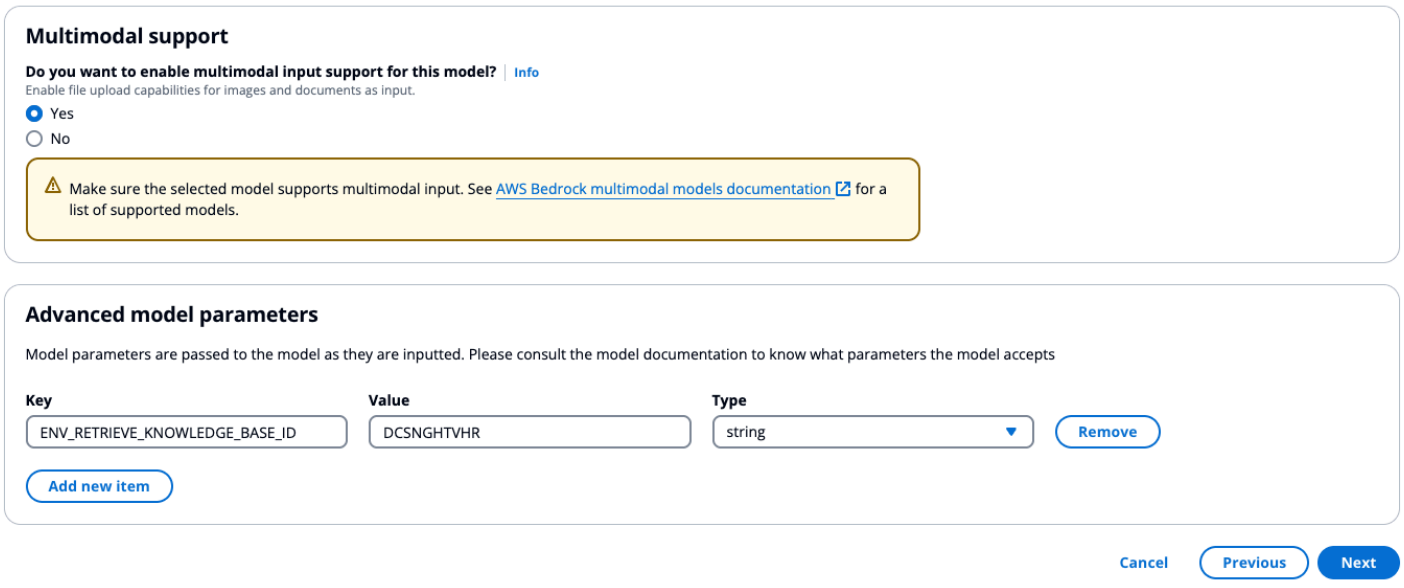
Opção 2: Parâmetros do modelo no assistente de implantação

Adicione variáveis de ambiente durante a etapa de seleção do modelo no assistente do Agent Builder usando a seção Parâmetros do modelo. As variáveis de ambiente que seguem a convenção de nomenclatura `ENV_<ALL_CAPS_TOOL_NAME>_<env_variable_name>` serão carregadas automaticamente em tempo de execução no ambiente de execução do agente como`<env_variable_name>`.

Por exemplo:

- `ENV_RETRIEVE_KNOWLEDGE_BASE_ID` se torna `KNOWLEDGE_BASE_ID`
- `ENV_RETRIEVE_MIN_SCORE` se torna `MIN_SCORE`

Seção de parâmetros avançados do modelo mostrando a configuração ENV_RETRIEVE_KNOWLEDGE_BASE_ID



Consulte a documentação ou o código-fonte da ferramenta específica para identificar as variáveis de ambiente necessárias. Para a ferramenta de recuperação, você pode encontrar opções de configuração no código-fonte.

Etapa 4: adicionar permissões do IAM

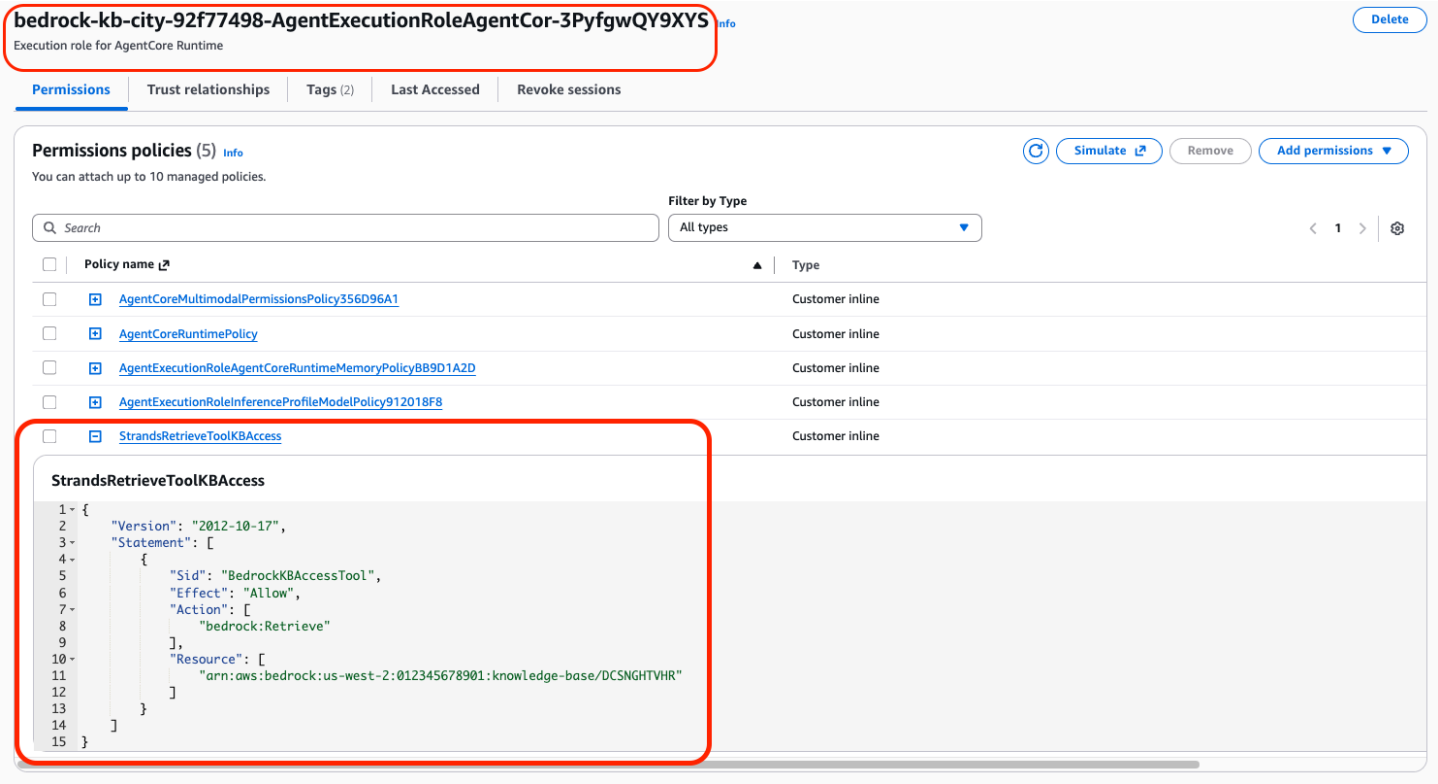
Adicione manualmente todas as permissões necessárias do IAM à sua função AgentCore de execução do Runtime para permitir que o agente use a ferramenta.

Por exemplo, para usar a ferramenta de recuperação com as bases de conhecimento Amazon Bedrock:

1. Navegue até o console do IAM em sua conta da AWS.
2. Localize a função AgentCore de execução do Runtime para seu agente.
3. Adicione a seguinte permissão:

```
{
  "Effect": "Allow",
  "Action": "bedrock:Retrieve",
  "Resource": "arn:aws:bedrock:region:account-id:knowledge-base/knowledge-base-id"
}
```

Console do IAM mostrando a StrandsRetrieveTool KBAccess política anexada à função AgentCore de execução do Runtime

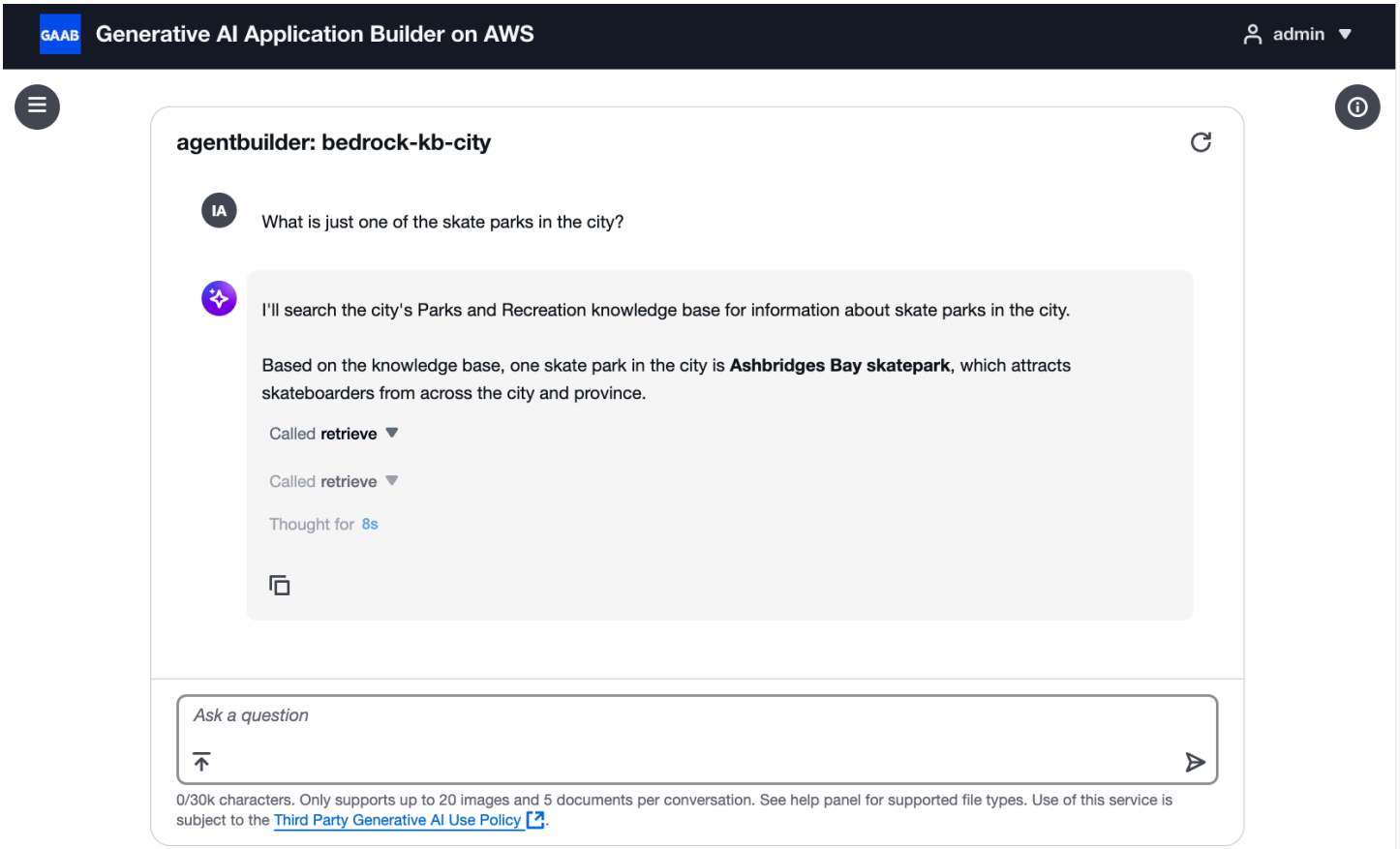


As permissões específicas necessárias variam de acordo com a ferramenta. Consulte a documentação da ferramenta e a documentação do serviço da AWS para determinar as permissões apropriadas do IAM.

Etapa 5: testar o agente

Depois de concluir as etapas de configuração, teste seu agente para verificar se a ferramenta está funcionando corretamente. Você deve ver as invocações da ferramenta nos registros de execução e nas respostas do agente.

Agente usando com sucesso a ferramenta de recuperação para responder a uma pergunta sobre parques de skate



> **Note**
>
> Para obter uma lista completa das ferramentas Strands disponíveis e seus recursos, consulte a documentação das Strands Community Tools.

Expandindo as bases de conhecimento suportadas e os tipos de memória de conversação

Para adicionar suas implementações de memória de conversação ou base de conhecimento, adicione as implementações necessárias na `shared` pasta e, em seguida, edite as fábricas e as enumerações apropriadas para criar uma instância dessas classes.

Quando você fornece a configuração do LLM, que é armazenada dentro do repositório de parâmetros, a memória de conversação e a base de conhecimento apropriadas serão criadas para seu LLM. Por exemplo, quando o `ConversationMemoryType` é especificado como DynamoDB, uma instância `DynamoDBChatMessageHistory` de (disponível `shared_components/memory/ddb_enhanced_message_history.py` no interior) é criada. Quando o `KnowledgeBaseType`

é especificado como Amazon Kendra, uma instância `KendraKnowledgeBase` de (disponível `shared_components/knowledge/kendra_knowledge_base.py` no interior) é criada.

Criando e implantando as alterações de código

Crie o programa com o `npm run build` comando. Depois que todos os erros forem resolvidos, execute `cdk synth` para gerar os arquivos de modelo e todos os ativos do Lambda.

1. Você pode usar o `0/stage-assets.sh` script para colocar manualmente todos os ativos gerados no intervalo de preparação da sua conta.
2. Use o comando a seguir para implantar ou atualizar a plataforma:

   ```
   cdk deploy DeploymentPlatformStack --parameters AdminUserEmail='admin-
   email@amazon.com'
   ```

   Quaisquer CloudFormation parâmetros adicionais da AWS também devem ser fornecidos junto com o AdminUserEmailparâmetro.

Guia de personalização

Gerenciando o grupo de usuários do Cognito

Quando o painel de implantação é implantado, um grupo de usuários do Amazon Cognito e um usuário administrador são criados para fornecer autenticação para o aplicativo. Esse grupo de usuários é compartilhado no painel de implantação e em todos os casos de uso. O usuário administrador criado na implantação do painel recebe automaticamente acesso a todos os casos de uso implantados usando o painel. Esse mecanismo é fornecido por meio de grupos de grupos de usuários do Amazon Cognito.

Quando um caso de uso é implantado a partir do painel, se um e-mail for fornecido, um usuário será criado no grupo de usuários compartilhado, junto com um grupo de usuários nomeado para o caso de uso específico. O usuário recém-criado é então adicionado ao grupo, concedendo ao usuário acesso ao caso de uso.

Se você deseja adicionar um usuário adicional a um determinado caso de uso, isso pode ser feito criando um usuário no grupo de usuários do Cognito e adicionando-o aos grupos correspondentes aos casos de uso aos quais você deseja que o usuário tenha acesso. Para obter um step-by-step guia, consulte Criação de um novo usuário no AWS Management Console.

Da mesma forma, se você quiser criar usuários administrativos adicionais, deverá criar um novo usuário e adicioná-lo ao grupo Administrador no grupo de usuários.

Os nomes de usuário são criados pegando a parte do e-mail fornecido antes do e anexando o `@` UUID do caso de uso gerado (ou `-admin` no caso do usuário administrador).

Na guia Grupos, você pode ver que um grupo de administradores e um grupo para cada caso de uso foram criados automaticamente usando o nome do caso de uso (conforme fornecido no assistente) e o UUID do caso de uso.

Referência de API

Esta seção fornece referências de API para a solução.

Painel de implantação

API REST	Método HTTP	Funcionalidade	Chamadores autorizados
`/deployments`	`GET`	Obtenha todas as implantações.	Token JWT autentica do pelo Amazon Cognito
`/deployments`	`POST`	Cria uma nova implantação de caso de uso.	Token JWT autentica do pelo Amazon Cognito
`/deployments/{useCaseId}`	`GET`	Obtém detalhes de implantação para uma única implantação.	Token JWT autentica do pelo Amazon Cognito
`/deployments/{useCaseId}`	`PATCH`	Atualiza uma determinada implantação.	Token JWT autentica do pelo Amazon Cognito
`/deployments/{useCaseId}`	`DELETE`	Exclui uma determina da implantação.	Token JWT autentica do pelo Amazon Cognito

API REST	Método HTTP	Funcionalidade	Chamadores autorizados
`/model-info/use-case-types`	`GET`	Obtém os tipos de casos de uso disponíveis para a implantação	Token JWT autenticado pelo Amazon Cognito
`/model-info/{useCaseType}/providers`	`GET`	Obtém os provedores de modelos disponíveis para o determinado tipo de caso de uso	Token JWT autenticado pelo Amazon Cognito
`/model-info/{useCaseType}/{providerName}`	`GET`	Obtém IDs os modelos disponíveis para um determinado provedor e tipo de caso de uso	Token JWT autenticado pelo Amazon Cognito
`/model-info/{useCaseType}/{providerName}/{modelId}`	`GET`	Obtém as informações sobre o modelo fornecido, incluindo os parâmetros padrão.	Token JWT autenticado pelo Amazon Cognito

> **Note**
>
> Os arquivos OpenAPI e Swagger também podem ser exportados do API Gateway para facilitar a integração com a API. Consulte Exportar uma API REST do API Gateway.

Cargas úteis POST e PATCH

Veja abaixo um exemplo de uma carga POST para o `/deployments` endpoint, que criará um novo caso de uso.

```
{
 "UseCaseName": "usecase1",
```

```
"UseCaseDescription": "Description of the use case to be deployed. For display
purposes", // optional
"DefaultUserEmail": "placeholder@example.com", // optional, if not provided, the
Cognito Group and User will not be created
"DeployUI": true, // optional
"VpcParams": {
"VpcEnabled": true,
"CreateNewVpc": false,
// provide these if not creating new vpc
"ExistingVpcId": "vpc-id",
"ExistingPrivateSubnetIds": ["subnet-1", "subnet-2"],
"ExistingSecurityGroupIds": ["sg-1", "sg-2"]
},
"ConversationMemoryParams": {
"ConversationMemoryType": "DynamoDB",
"HumanPrefix": "user", // optional
"AiPrefix": "ai", // optional
"ChatHistoryLength": 10 // optional
},
"KnowledgeBaseParams": {
"KnowledgeBaseType": "Bedrock",
// one of the following based on selected provider
"BedrockKnowledgeBaseParams": {
"BedrockKnowledgeBaseId": "my-bedrock-kb",
"RetrievalFilter": {}, // optional
"OverrideSearchType": "HYBRID" // optional
},
"KendraKnowledgeBaseParams": {
"AttributeFilter": {}, // optional
"RoleBasedAccessControlEnabled": true, // optional
"ExistingKendraIndexId": "12345678-abcd-1234-abcd-1234567890ab",
// provide the following in place of ExistingKendraIndexId if you want the solution to
deploy an index for you
"KendraIndexName": "index",
"QueryCapacityUnits": 1, // optional
"StorageCapacityUnits": 1, // optional
"KendraIndexEdition": "DEVELOPER" // optional
},
"NoDocsFoundResponse": "Sorry, I couldn't find any relevant information for your
query.", // optional
"NumberOfDocs": 3, // optional
"ScoreThreshold": 0.7, // optional
"ReturnSourceDocs": true // optional
},
```

```
 "LlmParams": {
 "ModelProvider": "Bedrock | SAGEMAKER",
 // one of the following based on selected provider
 "BedrockLlmParams": {
 "ModelId": "model-id", // use this for on demand models. Can't use with ModelArn
 "ModelArn": "model-arn", // use this for provisioned/custom models. Can't use with
 ModelId,
 "InferenceProfileId": "profile-id"
 "GuardrailIdentifier": "arn:aws:bedrock:us-east-1:123456789012:guardrail/my-
guardrail", // optional
 "GuardrailVersion": "1" // optional. Required if GuardrailIdentifier provided.
 },
 "SageMakerLlmParams": {
 "EndpointName": "some-endpoint",
 "ModelInputPayloadSchema": {},
 "ModelOutputJSONPath": "$."
 },
 // optional. Passes on arbitrary params to the underlying LLM.
 "ModelParams": {
 "param1": {
 "Value": "value1",
 "Type": "string"
 },
 "param2": {
 "Value": 1,
 "Type": "integer"
 }
 },
 // optional
 "PromptParams": {
 "PromptTemplate": "some template",
 "UserPromptEditingEnabled": true,
 "MaxPromptTemplateLength": 1000,
 "MaxInputTextLength": 1000,
 "DisambiguationPromptTemplate": "some disambiguation template",
 "DisambiguationEnabled": true
 },
 "Temperature": 1.0, // optional
 "Streaming": true, // optional
 "RAGEnabled": true, // optional. Must be true if providing KnowledgeBaseParams above.
 "Verbose": false // optional
 },
 "AgentParams": {
 "AgentType": "Bedrock",
```

```
 "BedrockAgentParams": {
 "AgentId": "agent-id",
 "AgentAliasId": "alias-id",
 "EnableTrace": true
 }
 },
 // optional
 "AuthenticationParams": {
 "AuthenticationProvider": "Cognito",
 "CognitoParams": {
 "ExistingUserPoolId": "user-pool-id",
 "ExistingUserPoolClientId": "client-id" // optional. If not provided, the solution
 will create a client for you in the provided pool
 }
 }

}
```

Para atualizações, a estrutura é a mesma acima, com algumas ressalvas:

- O nome do caso de uso não pode ser alterado
- Um caso de uso só pode alterar grupos de segurança e sub-redes depois de implantado em uma VPC. O VPC em si não pode ser alterado.
- Se um índice Kendra foi criado para você como uma base de conhecimento, você não pode alterar a configuração desse índice (por exemplo,,) KendraIndexName QueryCapacityUnits

Caso de uso compartilhado APIs

Os seguintes endpoints da API REST estão disponíveis para casos de uso do Text e do Bedrock Agent:

API REST	Método HTTP	Funcionalidade	Chamadores autorizados
`/details/{useCaseConfigKey}`	`GET`	Obtém detalhes de configuração para um caso de uso específico.	Token JWT autenticado pelo Amazon Cognito

WebSocket API	Funcionalidade	Chamadores autorizados
`/$connect`	Inicie a WebSocket conexão e autentique o usuário.	Token JWT autenticado pelo Amazon Cognito
`/$disconnect`	Endpoint chamado quando uma WebSocket conexão foi desconectada.	Token JWT autenticado pelo Amazon Cognito

API de detalhes do caso de uso

O endpoint de detalhes da API recupera informações sobre um caso de uso específico:

```
GET /details/{useCaseConfigKey}
```

Esse endpoint retorna os detalhes da configuração de um caso de uso específico, incluindo parâmetros do modelo, configurações da base de conhecimento e outras informações de implantação. Ele exige um token JWT autenticado pelo Amazon Cognito para autorização.

Caso de uso de texto

WebSocket API	Funcionalidade	Chamadores autorizados
`/sendMessage`	Envia a mensagem de bate-papo do usuário ao WebSocket para processam ento com a experiência LLM configurada.	Token JWT autenticado pelo Amazon Cognito

API REST	Método HTTP	Funcionalidade	Chamadores autorizados
`/feedback/{useCaseId}`	`POST`	Envia feedback do usuário sobre um	Token JWT autentica do pelo Amazon Cognito

API REST	Método HTTP	Funcionalidade	Chamadores autorizados
		caso de uso específico.	

Cargas úteis de envio de mensagens

Se você estiver se integrando diretamente à `/sendMessage` API, deverá seguir os seguintes formatos de carga útil de solicitação e resposta.

Solicitar carga

```
{
 "action": "sendMessage",
 "question": "the message to send to the api",
 "conversationId": "", // If not provided, a new conversation will be created, with the
 conversationId returned in the response. All subsequent messages in that conversation
 (where history is retained), should provide the conversationId there.
 "promptTemplate": "", // Optional. Overrides the configured prompt
 "authToken": "XXXX" // Optional. accessToken from cognito flow. Required for RAG with
 RBAC
}
```

Nome do parâmetro	Tipo	Description
ação	`String`	Atualmente, oferecemos suporte apenas à ação “SendMessage” no WebSocket
pergunta	`String`	A entrada do usuário a ser enviada ao LLM
ID da conversa	`String`	Um UUID identificando a conversa. Se não for fornecida, uma nova conversa será criada, com o ID da conversa retornado na resposta. Todas

Nome do parâmetro	Tipo	Description
		as mensagens subsequentes dessa conversa (onde você deseja que sejam retidas) devem fornecer o ID da conversa lá. history/context
Modelo de prompt	`String`[Opcional]	Substitui o modelo de prompt dessa mensagem. Se estiver vazio ou não for fornecido , usará como padrão o prompt definido no momento da implantação. Deve ter os espaços reservados adequados especificados para a configuração especific ada (ou seja, {history} e {input} para implantações do Sagemaker AI que não sejam do RAG, com a adição de {context} se estiver usando o RAG para todas as implantaç ões.

Nome do parâmetro	Tipo	Description
Token de autenticação	`String`[Opcional]	AccessToken foi obtido do fluxo de autenticação cognito. Isso é necessário ao invocar um endpoint de websocket de chat configura do para RAG com controle de acesso baseado em função (RBAC). A lista de declaraçõ es cognito:groups nesse token JWT é usada para controlar o acesso aos documento s no índice Kendra. Esse parâmetro não é necessário para casos de uso que não sejam do RAG. Também não é necessário para casos de uso do RAG com o RBAC desativado.

Cargas úteis de resposta

Resposta à pergunta

A WebSocket API responderá com 1 (se o streaming estiver desativado) ou vários (se o streaming estiver ativado) objetos JSON estruturados da seguinte forma para cada consulta.

```
{
 "data": "some data",
 "conversationId": "id",
}
```

Nome do parâmetro	Tipo	Description
data	`String`	Uma parte da resposta do LLM, se o streaming estiver

Nome do parâmetro	Tipo	Description
		ativado, ou a resposta inteira. Se estiver usando streaming , uma resposta desse formato com o conteúdo dos dados sendo END_CONVERSATION será enviada para indicar o final da resposta a uma única pergunta.
ID da conversa	`String`	O ID da conversa à qual essa resposta do SourceDocument pertence.

Resposta do documento de origem

Se você configurou seu caso de uso do RAG para retornar documentos de origem, você também receberá a seguinte carga útil no final de cada resposta para cada documento de origem usado para criar a resposta.

```
{
 "sourceDocument": {
 "excerpt": "some excerpt from the",
 "location": "s3://fake-bucket/test.txt",
 "score": 0.500,
 "document_title": null,
 "document_id": null,
 "additional_attributes": null
 },
 "conversationId": "some-id"
}
```

Nome do parâmetro	Tipo	Description
trecho	`String`	Um trecho do documento fonte.

Nome do parâmetro	Tipo	Description
location	`String`	Localização do documento de origem. Isso dependerá das fontes de dados usadas e do tipo de base de conhecimento, mas pode ser algo como s3 URIs ou sites.
pontuar	`Number` | `String`	A confiança de que o documento corresponde à pergunta feita. Isso será um float de 0 a 1 para Bedrock e uma string (por exemplo, HIGH, LOW, etc.) para Kendra.
título_documento	`String`	Título do documento fonte retornado. Disponível somente ao usar Kendra.
id_do_documento	`String`	ID do documento fonte retornado. Disponível somente ao usar Kendra.
atributos_adicionais	`String`	Esse campo conterá todos os atributos adicionais no documento, conforme personalizados em sua base de conhecimento no momento da ingestão.
ID da conversa	`String`	O ID da conversa à qual essa resposta do SourceDocument pertence.

Carga útil da API de feedback

Abaixo está um exemplo de uma carga POST para o `/feedback/{useCaseId}` endpoint, que enviará feedback do usuário para um caso de uso específico:

```
{
  "useCaseRecordKey": "12345678-12345678",
  "conversationId": "12345678-1234-1234-1234-123456789012",
  "messageId": "12345678-1234-1234-1234-123456789012",
  "feedback": "positive",
  "feedbackReason": ["accurate", "helpful"],
  "comment": "This response was very helpful.",
  "rephrasedQuery": "What are the key features of Amazon Bedrock?",
  "sourceDocuments": [
    "s3://bucket-name/document1.pdf",
    "s3://bucket-name/document2.pdf"
  ]
}
```

Caso de uso do Bedrock Agent

WebSocket API	Funcionalidade	Chamadores autorizados
`/invokeAgent`	Envia a mensagem do usuário ao WebSocket para processamento com o agente configurado.	Token JWT autenticado pelo Amazon Cognito

Cargas úteis do InvokeAgent

Se você estiver se integrando diretamente ao`/invokeAgent API`, deverá seguir os seguintes formatos de carga útil de solicitação e resposta.

Carga da solicitação

```
{
"action": "invokeAgent",
"inputText": "User query to the agent",
"conversationId": "", // Optional. Empty conversationId implies a new conversation.
 When not provided, a new conversationId will be created and returned with the
 response. All subsequent messages in the same conversation should provide the same
 conversationId (i.e. chat memory/history is maintained).
```

```
"authToken": "XXXX" // Optional. accessToken from cognito flow. If provided, it needs
 to be a valid JWT token associated with the user
}
```

Nome do parâmetro	Tipo	Description
ação	`String`	Apoiamos apenas a `invokeAgent` ação no WebSocket.
Texto de entrada	`String`	A entrada do usuário a ser enviada ao LLM.
ID da conversa	`String[Optional]`	Um UUID que identifica a conversa de forma exclusiva . Se você não fornecer esse valor, a solução criará uma nova conversa e retornará o ID da conversa na resposta. Todas as mensagens subsequentes nessa conversa (nas quais você deseja reter o histórico e o contexto) fornecem o ID da conversa lá.
Token de autenticação	`String[Optional]`	O AccessToken foi obtido do fluxo de autenticação do Amazon Cognito. Esse parâmetro não é obrigatório. Se você o fornecer, o token JWT será validado. Isso ajuda a facilitar a extensão dessa solução.

Cargas úteis de resposta

Resposta à pergunta

A WebSocket API responderá com um (se o streaming estiver desativado) ou vários (se o streaming estiver ativado) objetos JSON estruturados da seguinte forma para cada consulta.

```
{
 "data" "some data",
 "conversationId": "id",
 }
```

Nome do parâmetro	Tipo	Description
data	`String`	A resposta da invocação do agente.
ID da conversa	`String`	O ID da conversa.

Referência

Esta seção inclui informações sobre a coleta de dados dessa solução, indicadores para recursos relacionados e uma lista dos criadores que contribuíram para essa solução.

Provedores de LLM compatíveis

A solução pode se integrar aos seguintes fornecedores de LLM:

1. Amazon Bedrock
   - Documentação: https://aws.amazon.com/bedrock/
   - Modelos compatíveis:
     - Amazon
       - Nova Lite
       - Nova Micro
       - Nova Pro
     - AI21 Laboratórios
       - Jamba 1.5 Mini
       - Jamba 1.5 Large
     - Anthropic
       - Claude v3 Haicai
       - Soneto Claude v3.5
       - Claude v3.7 Sonnet (por meio do uso de perfis de inferência)
     - Cohere
       - Comando R
       - Command R+
     - Busca profunda
       - Deepseek-R1 (por meio do uso de perfis de inferência)
     - Meta
       - Llama 3
       - Llama 3.2 (por meio do uso de perfis de inferência)
     - Mistral AI

        - Mistral 7B Instruct
        - Instrução Mistral 8x7B
    - Inferência entre regiões
        - Capacidade de usar perfis de inferência definidos na mesma região do painel de implantação
2. SageMaker IA da Amazon
    - Documentação: https://aws.amazon.com/sagemaker/
    - Modelos compatíveis: modelos de texto para texto

Para obter os parâmetros mais recentes do modelo, as melhores práticas e os usos recomendados, consulte a documentação dos fornecedores do modelo.

Coleta de dados

Essa solução envia métricas operacionais para a AWS (os “Dados”) sobre o uso dessa solução. Usamos esses dados para entender melhor como os clientes usam essa solução e os serviços e produtos relacionados. A coleta desses dados pela AWS está sujeita ao Aviso de Privacidade da AWS.

Colaboradores

- Tarek Abdunabi
- Maid Arbash
- George Bearden
- Mukit Bin Momin
- Michael Connor
- Johnny Duval
- Nihit Kasabwala
- Ahern Knox
- Simon Krol
- Michael Lin
- Tim Mekari

- Ibrahim Mohamed
- Omar Radwan Mohsen
- James Nixon
- Dekshitha Ravikumar
- Jae Shim
- Ajay Swamy
- Mohammed Taha
- Reinicie Takkar
- Dimitri Tchikatilov
- Coroa de Jason
- Kamyar Ziabari

Revisões

Data de publicação: outubro de 2023 (última atualização: janeiro de 2025)

Verifique o arquivo CHANGELOG.md no GitHub repositório para ver todas as alterações e atualizações notáveis do software. O changelog fornece um registro claro das melhorias e correções referentes a cada versão.

Notices

Os clientes são responsáveis por fazer uma avaliação independente das informações contidas neste documento. Este documento: (a) serve apenas para fins informativos, (b) representa as práticas e ofertas atuais de produtos da AWS, que estão sujeitas a alterações sem aviso prévio, e (c) não cria nenhum compromisso ou garantia por parte da AWS e de seus afiliados, fornecedores ou licenciadores. Os produtos ou serviços da AWS são fornecidos “no estado em que se encontram”, sem garantias, representações ou condições de qualquer tipo, expressas ou implícitas. As responsabilidades e as obrigações da AWS para com os clientes são controladas por contratos da AWS, e este documento não faz parte nem modifica nenhum contrato entre a AWS e seus clientes.

O Generative AI Application Builder na AWS é licenciado sob os termos da Licença Apache Versão 2.0.

Important

O Generative AI Application Builder na AWS permite que você crie e implante aplicativos de inteligência artificial generativa na AWS usando o modelo de IA generativa de sua escolha, incluindo modelos de IA generativa de terceiros que você pode escolher usar e sobre os quais a AWS não possui ou sobre os quais não tem controle (“Modelos de IA generativa de terceiros”).

Seu uso dos Modelos de IA Generativos de Terceiros é regido pelos termos fornecidos pelos fornecedores de Modelos de IA Generativos de Terceiros quando você adquiriu sua licença para usá-los (por exemplo, seus termos de serviço, contrato de licença, política de uso aceitável e política de privacidade).

Você é responsável por garantir que o uso dos Modelos de IA Generativos de Terceiros esteja em conformidade com os termos que os regem e com quaisquer leis, regras, regulamentos, políticas ou padrões que se apliquem a você.

Você também é responsável por fazer sua própria avaliação independente dos modelos de IA generativa de terceiros que usa, incluindo seus resultados e como os fornecedores de modelos de IA generativa de terceiros usam quaisquer dados que possam ser transmitidos a eles com base em sua implantação. A AWS não faz nenhuma representação, garantia ou garantia em relação aos modelos de IA generativos de terceiros, que são “Conteúdo de terceiros” de acordo com seu contrato com a AWS. O Generative AI Application Builder na AWS é oferecido a você como “Conteúdo da AWS” de acordo com seu contrato com a AWS.

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.