Manual do usuário

Plataforma Claude na AWS

Plataforma Claude na AWS: Manual do usuário

Table of Contents

O que é a plataforma Claude na AWS?

Acesse os recursos completos da plataforma de Claude por meio da AWS com Anthropic-managed infraestrutura.

A Claude Platform na AWS oferece a experiência completa da plataforma Anthropic, incluindo a API de mensagens, habilidades do agente, execução de código e recursos beta, acessíveis por meio de sua conta da AWS. Ao contrário do Amazon Bedrock, onde a AWS opera a pilha de inferência, a Anthropic opera a Claude Platform na AWS. A AWS fornece a camada de autenticação (SigV4 ou chave de API), controle de IAM-based acesso e integração de faturamento por meio do AWS Marketplace.

> **Note**
>
> Os SDKs da Anthropic oferecem suporte à plataforma Claude na AWS. Para ver a disponibilidade do cliente por idioma, consulte a documentação dos SDKs do Anthropic Client.

Como funciona a integração da plataforma

Os modelos Claude são executados em Anthropic-managed infraestrutura. Essa é uma integração comercial para cobrança e acesso por meio da AWS. Tanto a Anthropic quanto a AWS atuam como processadores de dados independentes. Os Termos de Serviço da AWS regem a Plataforma Claude na AWS. Os Termos de Serviço Comerciais, o Adendo de Processamento de Dados e a Política de Uso da Anthropic também se aplicam.

Observe as seguintes características operacionais. Os dados podem não residir na AWS. A inferência pode ser direcionada para a nuvem primária da Anthropic. Os subserviços podem mudar sem aviso prévio. Defina o `inference_geo` parâmetro por solicitação para fixar a inferência em uma geografia específica. Consulte Residência de dados para obter detalhes.

A Claude Platform na AWS segue a mesma política de retenção de dados da API Claude primária. A retenção zero de dados (ZDR) está disponível mediante solicitação. Entre em contato com seu representante de conta da Anthropic para habilitá-lo para sua conta.

Características da plataforma Claude na AWS

A Claude Platform na AWS fornece os seguintes recursos:

- Same-day acesso a modelos e recursos — Os novos modelos e recursos de API da Claude estão disponíveis no mesmo dia em que são lançados na API original da Claude.
- Habilidades do agente — Use habilidades pré-criadas para geração de documentos (ExcelPowerPoint, Word, PDF) e crie habilidades personalizadas.
- Execução de código — Execute código na sandbox gerenciada da Anthropic.
- Pensamento estendido — Permita o pensamento estendido para tarefas complexas de raciocínio.
- Streaming e processamento em lote — use streaming SSE em tempo real ou envie solicitações em lote para cargas de trabalho de alto rendimento.
- Cache imediato — ferramentas de cache, solicitações do sistema e histórico de mensagens para reduzir a latência e o custo.
- API de arquivos — faça upload e referencie arquivos em todas as solicitações.
- Integração com o AWS IAM — controle o acesso com políticas padrão do IAM e autenticação SigV4.
- Faturamento unificado da AWS — Pague por meio de sua conta existente da AWS com medição baseada no consumo (Marketplace como back-end de cobrança).

Para ver a lista completa dos recursos compatíveis, consulte Suporte de recursos.

Como este guia está organizado

Este guia aborda os seguintes tópicos:

- Introdução — Configuração da conta, pré-requisitos, autenticação e instalação do SDK.
- Usando a API — Modelos disponíveis, solicitações e suporte a recursos.
- Gerenciando seu ambiente — residência de dados, espaços de trabalho, console Claude, limites de tarifas e cobrança.
- Monitoramento e operações — CloudTrail registro e rastreamento de ID de solicitação.
- Migração — migrando do Amazon Bedrock para a plataforma Claude na AWS.
- Segurança e controle de acesso — referência de políticas e ações do IAM.

Plataforma Claude na AWS versus Amazon Bedrock

Ambas as ofertas permitem que você use Claude por meio da AWS, mas elas diferem significativamente em arquitetura, superfície de API e disponibilidade de recursos.

	Plataforma Claude na AWS	Amazon Bedrock
Quem opera a pilha	Anthropic	AWS
Superfície da API	API de mensagens antrópicas () `/v1/messages`	API Bedrock (Converse/) InvokeModel
Disponibilidade de recursos	Same-day como Claude API	Depende do cronograma de integração com a AWS
Habilidades do agente	Available (Disponível)	Não disponível (requer execução de código)
Funcionalidades beta	Passe com `anthropic-beta` cabeçalhos (consulte Suporte de recursos)	Requer suporte da AWS
Autenticação	AWS IAM//SigV4 ou chave de API	AWS IAM/SigV4 ou token portador (somente SDKs para C#, Go e Java)
URL base	`aws-external-anthropic.{region}.api.aws`	`bedrock-runtime.{region}.amazonaws.com`
Cliente SDK	Platform-specific classe cliente (por exemplo, `AnthropicAWS` em Python), em beta	`AnthropicBedrock` /SDK Bedrock
Console	Claude Console (`platform.claude.com` , acesso por meio do AWS Console)	Consola Bedrock

	Plataforma Claude na AWS	Amazon Bedrock
Limites de tarifas e cotas	Gerenciado pela Anthropic	Gerenciado pela AWS
Processadores de dados	Anthropic e AWS	AWS

Se você precisar AWS-operated de Claude com a API Bedrock, consulte Amazon Bedrock.

> **Important**
>
> A Anthropic fornece a Claude Platform na AWS como uma oferta terceirizada. Ela não é coberta pelos programas de conformidade, certificações ou relatórios de auditoria padrão da AWS (como elegibilidade para SOC, ISO ou HIPAA). Os clientes são os únicos responsáveis por realizar sua própria diligência para garantir que essa oferta de terceiros atenda aos requisitos regulatórios, legais e de conformidade.

Quando escolher o Bedrock: escolha o Amazon Bedrock se sua organização exigir AWS-operated inferência, a AWS como o único processador de dados ou a cobertura dos programas de conformidade da AWS (incluindo elegibilidade para FedRAMP High, IL4, IL5, SOC, ISO e HIPAA). A Bedrock funciona inteiramente em AWS-controlled infraestrutura com a AWS como parte operacional.

Configurar a conta.

A configuração da plataforma Claude na AWS tem quatro fases: cadastre-se no console da AWS, configure sua organização antrópica, anote o ID do seu espaço de trabalho e faça login no console Claude.

> **Note**
>
> A inscrição por meio do AWS Console provisiona uma nova organização antrópica vinculada à sua conta da AWS. Essa organização é separada de todas as organizações existentes que sua empresa tem com a Anthropic, incluindo organizações da Claude Enterprise adquiridas por meio do AWS Marketplace. As chaves de API, os espaços de trabalho e as configurações do Claude Console de uma organização antrópica primária não são transferidas.
>
> Se você já tem uma oferta privada do Amazon Bedrock, entre em contato com seu executivo de contas da Anthropic ou da AWS antes de se inscrever para que seu desconto seja aplicado a partir de sua primeira solicitação. Os descontos não podem ser aplicados retroativamente ao uso incorrido antes que sua oferta privada seja aceita. Veja as ofertas privadas.

Etapa 1: cadastre-se no console da AWS

1. Abra o console da AWS e navegue até a plataforma Claude na página de serviço da AWS.
2. Escolha Inscrever-se.
3. Escolha Continuar.

A página mostra um banner Sign-up em andamento. Fique na página. Sign-up leva alguns minutos enquanto a AWS gerencia a assinatura do AWS Marketplace para você e, em seguida, redireciona você automaticamente.

Se sua organização tiver uma oferta privada da Anthropic, o console a pesquisará e solicitará que você a aceite no AWS Marketplace. Consulte Ofertas privadas para obter detalhes.

> **Note**
>
> Se você usa a Plataforma Claude na AWS, seu conteúdo (como solicitações e conclusões) é processado pela Anthropic fora da AWS. Consulte as políticas de uso de dados da Anthropic para obter detalhes sobre como o conteúdo e os metadados são processados e armazenados.

Etapa 2: configurar sua organização antrópica

Depois que a inscrição for concluída, você será redirecionado para. `platform.claude.com/partner-signup`

1. Insira o endereço de e-mail do proprietário da sua organização e escolha Começar.
2. Verifique se há um link de configuração na caixa de entrada do e-mail e siga-o. Se o seu navegador mostrar uma página Conectado como uma conta diferente, escolha Sair e continuar.
3. Preencha o formulário de detalhes da organização (nome da organização, tipo de entidade, país, uso pretendido) e escolha Configuração completa.

A conclusão da configuração cria sua organização antrópica. Os Termos de Serviço da AWS regem a Plataforma Claude na AWS. Os Termos de Serviço Comerciais, o Adendo de Processamento de Dados e a Política de Uso da Anthropic também se aplicam. A página de serviço do AWS Console agora mostra uma navegação à esquerda com Home, chaves de API, Quickstart e Workspaces.

Etapa 3: anote o ID do seu espaço de trabalho

Um espaço de trabalho padrão é provisionado automaticamente quando você se inscreve. Espaços de trabalho adicionais podem ser criados por região; consulte Espaços de trabalho para obter detalhes sobre vinculação de regiões, gerenciamento de espaços de trabalho e escopo de recursos do IAM.

Encontre o ID do espaço de trabalho em Espaços de trabalho no console da AWS, na plataforma Claude, na página de serviço da AWS ou no console Claude (consulte Uso do console Claude). Os IDs do espaço de trabalho usam o formato `wrkspc_` seguido por um identificador alfanumérico.

Etapa 4: Faça login no console Claude

O acesso ao console Claude é federado por meio do AWS IAM:

1. Assuma uma função do IAM com a `aws-external-anthropic:AssumeConsole` permissão. Veja as ações do IAM.
2. Na página de serviço Claude Platform na AWS, escolha Sign in. O AWS Console emite um JWT e redireciona você para o. `platform.claude.com`
3. No primeiro login, você será solicitado a fornecer um endereço de e-mail. Insira seu e-mail de trabalho. A plataforma provisiona seu usuário do Claude Console em tempo hábil.

Quando você faz login por meio do Console da AWS, o Console Claude define o escopo da sua plataforma Claude na organização da AWS. Um indicador Conta gerenciada pela AWS aparece na barra lateral do Claude Console.

Solução de problemas na configuração da conta

- "Sign-up falhou: Falha ao ativar OutboundWebIdentityFederation “: Se você ver esse banner vermelho no primeiro envio, escolha Continuar novamente. A ativação do IAM pode levar alguns minutos para se propagar.
- Nenhum indicador de progresso durante a inscrição: Sign-up leva alguns minutos. A página mostra um banner estático Sign-up em andamento sem uma barra de progresso enquanto a AWS provisiona sua conta.
- “Conectado como uma conta diferente” depois de seguir o link de configuração: Escolha Sair e continuar. A página reautentica você com o endereço de e-mail digitado.
- Mensagem “Não encontrada” durante o login: essa mensagem pode aparecer brevemente durante o redirecionamento. Você pode descartá-lo.
- A página de uso não mostra dados após sua primeira chamada de API: os dados de uso podem levar alguns minutos para aparecer no Claude Console.

Pré-requisitos

Antes de fazer chamadas de API, verifique se você tem:

1. Uma conta ativa da AWS com uma assinatura da Claude Platform na AWS (consulte Configurar sua conta).
2. A AWS CLI instalada e configurada.
3. Federação de identidade da web de saída ativada em sua conta da AWS (uma etapa de configuração única; consulte Habilitar federação de identidade da web de saída).
4. Seu ID do espaço de trabalho (consulte Obter seu ID do espaço de trabalho).

Habilitar federação de identidades web de saída

A plataforma Claude no gateway da AWS chama o `sts:GetWebIdentityToken` lado do servidor para criar um JWT que ele encaminha para a Anthropic. Esse recurso de STS está desativado por padrão em todas as contas da AWS. Ative-o uma vez por conta:

```
aws iam enable-outbound-web-identity-federation
```

Se a resposta for`[ERROR] (FeatureEnabled) … already enabled` , a configuração da sua conta já está ativada e você pode seguir em frente. Verifique e recupere o URL do emissor da sua conta:

```
aws iam get-outbound-web-identity-federation-info
```

> **Warning**
>
> Sem essa etapa, todas as solicitações retornam`"Outbound web identity federation is disabled for your account"`. Esse é o erro de configuração mais comum.

Obtenha seu ID do espaço de trabalho

Um espaço de trabalho padrão é provisionado automaticamente em cada região quando você se inscreve (consulte Configurar sua conta). Os espaços de trabalho estão vinculados a uma única

região da AWS. Você pode encontrar o ID do espaço de trabalho no Console Claude, em Espaços de trabalho, ou na seção Espaços de trabalho da página de serviço do Console AWS. Consulte Espaços de trabalho para vinculação de regiões e escopo de recursos do IAM.

Defina as variáveis de `AWS_REGION` ambiente `ANTHROPIC_AWS_WORKSPACE_ID` e para que os clientes do SDK as leiam automaticamente:

```
export ANTHROPIC_AWS_WORKSPACE_ID='wrkspc_01AbCdEf23GhIj'
export AWS_REGION='us-west-2'
```

A região é obrigatória. O cliente SDK é lançado se nenhuma região estiver definida. Passe`aws_region`/`awsRegion`para o construtor ou defina `AWS_REGION` (ou`AWS_DEFAULT_REGION`). Todas as regiões comerciais da AWS são suportadas; as regiões de inscrição exigem que você primeiro inscreva a conta na região.

Autenticação

A plataforma Claude na AWS oferece suporte a dois métodos de autenticação: AWS IAM com assinatura de solicitação SigV4 (primária) e autenticação de chave de API. Ambos usam o mesmo URL base e formato de solicitação.

Os Anthropic SDKs (consulte Instalar um SDK) implementam o fluxo de autenticação e a resolução de credenciais descritos abaixo. Se você não estiver usando um SDK antrópico, deverá criar SigV4-signed solicitações (ou apresentar sua chave de API como um token portador) no endpoint regional. `https://aws-external-anthropic.<region>.api.aws` Para a configuração SDK-specific do cliente e a ordem autorizada de resolução de credenciais, consulte o guia de configuração Claude on AWS na documentação da Anthropic.

Autenticação SigV4

O SigV4 é o caminho nativo corporativo e se integra às suas políticas, funções e auditorias existentes do AWS IAM. Configure as credenciais da AWS usando qualquer método suportado pela cadeia de fornecedores de credenciais padrão da AWS:

- Variáveis de ambiente (`AWS_ACCESS_KEY_ID`,`AWS_SECRET_ACCESS_KEY`,`AWS_SESSION_TOKEN`)
- Arquivo de credenciais compartilhado () `~/.aws/credentials`
- Arquivo de configuração compartilhado (`~/.aws/config`) incluindo SSO e `credential_process`
- Identidade da Web (`AWS_WEB_IDENTITY_TOKEN_FILE`e`AWS_ROLE_ARN`) para IRSA e Actions GitHub
- Credenciais de contêiner do ECS
- Serviço de metadados de instância do EC2 (IMDS)

Para verificar se o Anthropic SDK está coletando suas credenciais e indo para o endpoint regional correto, faça uma solicitação de teste seguindo os exemplos em Fazer solicitações. Uma resposta bem-sucedida confirma que as credenciais, a região, o URL base e o ID do espaço de trabalho estão todos configurados corretamente.

Autenticação de chave de API

Para caminhos de integração mais simples (desenvolvimento local e scripts), você pode se autenticar com uma chave de API em vez de SigV4. Defina a variável de `ANTHROPIC_AWS_API_KEY` ambiente ou passe `apiKey` para o construtor do SDK. O Anthropic SDK envia a chave como um token portador para a plataforma Claude no endpoint da AWS; o IAM autoriza a solicitação por meio da `aws-external-anthropic:CallWithBearerToken` ação (consulte as políticas do IAM).

Gere chaves de API no console da AWS em Claude Platform na AWS → Chaves de API. Escolha Gerar uma chave e, em seguida, copie o valor da chave. Conceda a ação `aws-external-anthropic:CallWithBearerToken` do IAM aos diretores que devem ter permissão para usar a autenticação por chave de API.

> **Note**
>
> Somente as chaves de API criadas no console da AWS na plataforma Claude na AWS funcionam com esse serviço.
>
> - As chaves criadas no console Claude padrão para acesso primário à API não funcionam na plataforma Claude no endpoint da AWS.
> - As chaves de API do Amazon Bedrock também não funcionam — o Bedrock usa um endpoint, um fluxo de autenticação e um namespace do IAM separados.

Short-term Chaves de API

Nos casos em que você deseja a autenticação da chave de API, mas sem a vida útil de uma chave de longa duração, a Anthropic publica bibliotecas geradoras de tokens que geram chaves de API de curto prazo a partir de suas credenciais do AWS IAM. Os tokens têm como padrão uma vida útil de 12 horas e podem ser definidos para um espaço de trabalho específico e para a ação. `aws-external-anthropic:CallWithBearerToken` Instale o gerador para sua linguagem, troque as credenciais do IAM por uma chave de API e passe a chave para o Anthropic SDK.

- Python: -gerador para aws-external-anthropic-python aws/token
- JavaScript / TypeScript: aws/token-gerador para aws-external-anthropic-js
- Java: aws/token-gerador para aws-external-anthropic-java

Short-term As chaves de API ainda exigem que o diretor do IAM do chamador permaneça `aws-external-anthropic:CallWithBearerToken` no espaço de trabalho de destino. Eles expiram sozinhos e não precisam ser alternados ou revogados explicitamente.

Precedência de credenciais e resolução de região

A plataforma Claude do Anthropic SDK no cliente da AWS resolve as credenciais e a região usando uma ordem de precedência definida. Os nomes dos argumentos seguem as convenções de cada linguagem: CamelCase para PHP, snake_case TypeScript para Python e Ruby, para Go e padrões de propriedade ou construtor para PascalCase C# e Java.

Para obter a ordem de precedência autorizada, as variáveis de ambiente suportadas e os argumentos do construtor para cada SDK, consulte Claude on AWS setup na documentação da Anthropic. Em geral, os argumentos explícitos do construtor têm precedência sobre as variáveis de ambiente e têm `ANTHROPIC_AWS_API_KEY` precedência sobre a cadeia de fornecedores de credenciais padrão da AWS. A região é obrigatória; diferentemente de `AnthropicBedrock` (que se baseia em`us-east-1`), o cliente Claude on AWS lança se nenhuma região for fornecida pelo construtor ou por/. `AWS_REGION` `AWS_DEFAULT_REGION`

ID do Workspace

Cada solicitação do plano de dados deve incluir o ID do espaço de trabalho no `anthropic-workspace-id` cabeçalho. Os SDKs da Anthropic leem isso da variável de `ANTHROPIC_AWS_WORKSPACE_ID` ambiente por padrão, ou você pode passar`workspaceId`/`workspace_id`para o construtor do cliente. Se você usar o `Anthropic` cliente base (não o Claude-on-AWS cliente), passe o cabeçalho explicitamente. Consulte Como fazer solicitações de exemplos por idioma e Espaços de trabalho para saber como localizar o ID do seu espaço de trabalho.

Instalar um SDK do

Os SDKs de clientes da Anthropic oferecem suporte à Claude Platform na AWS. Todos os sete SDKs fornecem uma classe de cliente específica da plataforma; como alternativa, você pode configurar o cliente base com a URL base da plataforma Claude na AWS.

Python

```
pip install -U "anthropic[aws]"
```

> **Tip**
>
> No macOS com Homebrew Python ou outros ambientes `pip install` Python gerenciados externamente, pode falhar com um erro PEP 668. `externally-managed-environment` Crie e ative primeiro um ambiente virtual:`python3 -m venv .venv && source .venv/bin/activate`.

TypeScript

```
npm install @anthropic-ai/aws-sdk
```

C#

```
dotnet add package Anthropic.Aws
```

Go

```
go get github.com/anthropics/anthropic-sdk-go
```

Java

```
implementation("com.anthropic:anthropic-java-aws:2.27.0")
```

```
<dependency>
  <groupId>com.anthropic</groupId>
  <artifactId>anthropic-java-aws</artifactId>
```

```
  <version>2.27.0</version>
</dependency>
```

PHP

```
composer require anthropic-ai/sdk aws/aws-sdk-php
```

Ruby

```
gem install anthropic aws-sdk-core
```

> **Note**
>
> Os clientes do SDK da Claude Platform na AWS estão na versão beta.

Modelos disponíveis

A Claude Platform na AWS oferece o mesmo conjunto de modelos Claude que a API Claude original.

Para obter a lista atual de modelos, IDs de modelo, tamanhos de janela de contexto e recursos, consulte Visão geral dos modelos na documentação da Anthropic.

IDs do modelo

Os IDs de modelo na plataforma Claude na AWS são idênticos aos usados na API Claude primária (por exemplo,,`claude-opus-4-6`,`claude-sonnet-4-6`). `claude-haiku-4-5` Não há Bedrock-style ARNs ou `anthropic.` prefixos — use o ID do modelo exatamente como a Anthropic o publica.

Fazer solicitações

A Claude Platform na AWS usa a Anthropic Messages API (`/v1/messages`), a mesma superfície de API da plataforma primária Claude. As diferenças são a URL base, o método de autenticação e um `anthropic-workspace-id` cabeçalho obrigatório que identifica qual espaço de trabalho a solicitação tem como alvo.

Concha

```
curl "https://aws-external-anthropic.us-west-2.api.aws/v1/messages" \
  --aws-sigv4 "aws:amz:us-west-2:aws-external-anthropic" \
  --user "$AWS_ACCESS_KEY_ID:$AWS_SECRET_ACCESS_KEY" \
  -H "x-amz-security-token: $AWS_SESSION_TOKEN" \
  -H "content-type: application/json" \
  -H "anthropic-version: 2023-06-01" \
  -H "anthropic-workspace-id: $ANTHROPIC_AWS_WORKSPACE_ID" \
  -d '{
    "model": "claude-sonnet-4-6",
    "max_tokens": 1024,
    "messages": [
      {"role": "user", "content": "Hello!"}
    ]
  }'
```

Python

```
from anthropic import AnthropicAWS

client = AnthropicAWS()

message = client.messages.create(
    model="claude-sonnet-4-6",
    max_tokens=1024,
    messages=[{"role": "user", "content": "Hello!"}],
)
print(message)
```

TypeScript

```
import AnthropicAws from "@anthropic-ai/aws-sdk";
```

```
const client = new AnthropicAws();

const message = await client.messages.create({
  model: "claude-sonnet-4-6",
  max_tokens: 1024,
  messages: [{ role: "user", content: "Hello!" }]
});
console.log(message);
```

C#

```
using Anthropic;
using Anthropic.Aws;

var client = new AnthropicAwsClient();

var message = await client.Messages.Create(new()
{
    Model = "claude-sonnet-4-6",
    MaxTokens = 1024,
    Messages = [new() { Role = "user", Content = "Hello!" }]
});

Console.WriteLine(message);
```

Go

```
client, err := anthropicaws.NewClient(context.Background(),
 anthropicaws.ClientConfig{})
if err != nil {
    panic(err)
}

message, err := client.Messages.New(context.Background(), anthropic.MessageNewParams{
    Model:     anthropic.ModelClaudeSonnet4_6,
    MaxTokens: 1024,
    Messages: []anthropic.MessageParam{
        anthropic.NewUserMessage(anthropic.NewTextBlock("Hello!")),
    },
})
if err != nil {
    panic(err)
```

```
}

fmt.Println(message.Content[0].Text)
```

Java

```
import com.anthropic.aws.backends.AwsBackend;
import com.anthropic.client.AnthropicClient;
import com.anthropic.client.okhttp.AnthropicOkHttpClient;
import com.anthropic.models.messages.Message;
import com.anthropic.models.messages.MessageCreateParams;
import com.anthropic.models.messages.Model;

AnthropicClient client = AnthropicOkHttpClient.builder()
    .backend(AwsBackend.fromEnv())
    .build();

Message message = client.messages().create(
    MessageCreateParams.builder()
        .model(Model.CLAUDE_SONNET_4_6)
        .maxTokens(1024)
        .addUserMessage("Hello!")
        .build()
);

IO.println(message);
```

PHP

```
use Anthropic\Aws\Client;

$client = new Client();

$message = $client->messages->create(
    model: 'claude-sonnet-4-6',
    maxTokens: 1024,
    messages: [['role' => 'user', 'content' => 'Hello!']],
);

echo $message;
```

Ruby

```
require "anthropic"

client = Anthropic::AWSClient.new

message = client.messages.create(
  model: "claude-sonnet-4-6",
  max_tokens: 1024,
  messages: [{ role: "user", content: "Hello!" }]
)

puts message
```

O cliente lê `AWS_REGION` (ou`AWS_DEFAULT_REGION`) e `ANTHROPIC_AWS_WORKSPACE_ID` do ambiente. Você pode substituir passando`aws_region`/`awsRegion`ou`workspace_id`/`workspaceId`para o construtor. Tanto a ID da região quanto a do espaço de trabalho são obrigatórias; o construtor lança se alguma delas não puder ser resolvida a partir dessas fontes.

> **Note**
>
> O `x-amz-security-token` cabeçalho (curl) só é necessário para credenciais temporárias, como funções do IAM, SSO ou STS. Omita-o ao usar credenciais de usuário do IAM de longo prazo. Os clientes do SDK lidam com isso automaticamente com base na fonte da credencial.

O `--aws-sigv4` valor segue o formato`aws:amz:<region>:<service>`. O nome do serviço SigV4 é`aws-external-anthropic`, e a região deve corresponder à região na URL do seu endpoint. Uma incompatibilidade em qualquer uma delas produz um erro genérico de rejeição de assinatura em vez de um diagnóstico específico.

Usando o cliente antrópico básico

Se você preferir não adicionar a dependência do AWS SDK, você pode usar o cliente base`Anthropic`. Esse caminho usa os nomes das variáveis de ambiente vanilla do SDK em vez dos `ANTHROPIC_AWS_*` nomes que o cliente dedicado lê:

```
export ANTHROPIC_API_KEY='your-api-key'
export ANTHROPIC_BASE_URL='https://aws-external-anthropic.us-west-2.api.aws'
```

```
export ANTHROPIC_WORKSPACE_ID='your-workspace-id'
```

Os SDKs do Python e Go são lidos `ANTHROPIC_API_KEY` e `ANTHROPIC_BASE_URL` automaticamente; para as outras linguagens, passe-os para o construtor. TypeScript Em cada idioma, passe o ID do espaço de trabalho no `anthropic-workspace-id` cabeçalho.

Python

```
import os
from anthropic import Anthropic

client = Anthropic(
    # ANTHROPIC_API_KEY and ANTHROPIC_BASE_URL are read automatically
    default_headers={"anthropic-workspace-id": os.environ["ANTHROPIC_WORKSPACE_ID"]},
)

message = client.messages.create(
    model="claude-sonnet-4-6",
    max_tokens=1024,
    messages=[{"role": "user", "content": "Hello!"}],
)
print(message.content[0].text)
```

TypeScript

```
import Anthropic from "@anthropic-ai/sdk";

const client = new Anthropic({
  // ANTHROPIC_API_KEY and ANTHROPIC_BASE_URL are read automatically
  defaultHeaders: { "anthropic-workspace-id": process.env.ANTHROPIC_WORKSPACE_ID }
});

const message = await client.messages.create({
  model: "claude-sonnet-4-6",
  max_tokens: 1024,
  messages: [{ role: "user", content: "Hello!" }]
});
console.log(message.content);
```

Suporte a recursos

A Claude Platform na AWS usa diretamente a API Anthropic Messages. Você obtém paridade total de recursos com a API Claude original, exceto quando indicado em Recursos não disponíveis no momento:

- Recursos beta: passe o `anthropic-beta` cabeçalho padrão para acessar os recursos beta, assim como você faria com a API Claude.
- Habilidades do agente: use habilidades de agente pré-criadas e personalizadas com os mesmos `container.skills` parâmetros e cabeçalhos beta da API Claude. Todas as habilidades pré-criadas (ExcelPowerPoint, Word, PDF) funcionam imediatamente.
- Execução de código: execute código na sandbox gerenciada da Anthropic usando a ferramenta de execução de código.
- Uso da ferramenta: O uso do computador e todos os outros recursos de uso da ferramenta estão disponíveis.
- Pensamento estendido: habilite o pensamento estendido com os mesmos parâmetros da API Claude.
- Streaming: suporte completo de streaming SSE para respostas em tempo real.
- Processamento em lote: envie solicitações em lote para cargas de trabalho de alto rendimento.
- Cache imediato: ferramentas de cache, solicitações do sistema e histórico de mensagens para reduzir a latência e o custo. Todos os recursos de cache imediato (TTL de 5 minutos, TTL de 1 hora e cache automático) estão disponíveis.
- API de arquivos: faça upload e referencie arquivos em todas as solicitações.
- Tags de espaço de trabalho com integração do IAM: os espaços de trabalho podem ser marcados e as tags fluem para o IAM para controle de acesso baseado em atributos e para os relatórios de uso e custos da AWS para alocação de custos. A marcação do espaço de trabalho é um recurso de GA na plataforma Claude na AWS (é um recurso beta na API original da Claude).

Claude Managed Agents

Os agentes gerenciados da Claude estão disponíveis na plataforma Claude na AWS. Agentes, sessões, ambientes, cofres de credenciais e armazenamentos de memória são recursos IAM de primeira classe. Consulte as ações do IAM para a referência da ação e Como usar o Claude Console

para as páginas correspondentes. O Anthropic Agent SDK funciona com a Claude Platform na AWS sem integração adicional. Aponte para a URL base da plataforma Claude na AWS e autentique-a com SigV4 ou uma chave de API.

As sessões autônomas na plataforma Claude na AWS exigem reautenticação a cada 6 horas. Long-running as execuções do agente devem atualizar suas credenciais SigV4 ou chave de API nessa janela, ou a sessão será encerrada.

No momento, os seguintes recursos de Claude Managed Agents não estão disponíveis na plataforma Claude na AWS:

- Resultados: o rastreamento de resultados para sessões de agentes não está disponível.
- Multi-agent sessões: sessões com vários agentes interagindo não estão disponíveis.

Compliance

> **Important**
>
> A Anthropic fornece esse serviço como uma oferta de terceiros. Ela não é coberta pelos programas de conformidade, certificações ou relatórios de auditoria padrão da AWS (como elegibilidade para SOC, ISO ou HIPAA). Os clientes são os únicos responsáveis por realizar sua própria diligência para garantir que essa oferta de terceiros atenda aos requisitos regulatórios, legais e de conformidade. Antes de processar dados confidenciais ou regulamentados, você deve revisar a documentação oficial de conformidade da Anthropic por meio do Anthropic Trust Center. Consulte os Termos de Serviço da AWS para obter mais informações.

Como a associação ao espaço de trabalho é modelada

Na API Claude primária, o acesso é regido pela associação de usuários a espaços de trabalho — um usuário é adicionado a um espaço de trabalho e herda os direitos de acesso do espaço de trabalho. A Claude Platform na AWS substitui esse modelo pela autorização do IAM: não há uma lista de usuários por espaço de trabalho. Em vez disso, os diretores do IAM (usuários ou funções) mantêm políticas que concedem `aws-external-anthropic` ações contra ARNs específicos do espaço de trabalho. A avaliação da política do IAM determina o que cada diretor pode fazer em cada espaço de trabalho.

Conceda e revogue o acesso ao espaço de trabalho modificando as políticas do IAM (SCPs, limites de permissão, políticas anexadas à identidade ou condições em nível de recurso) em vez de gerenciar a associação por espaço de trabalho no Claude Console. Consulte as políticas do IAM para ver exemplos.

Recursos não disponíveis no momento

No momento, os seguintes recursos não estão disponíveis na plataforma Claude na AWS:

- Preparação para a HIPAA: o HIPAA-ready programa da Anthropic não está disponível na plataforma Claude na AWS. Em vez disso, clientes com requisitos da HIPAA devem avaliar Claude no Amazon Bedrock.
- Níveis de serviço além do Standard e do Batch: o Priority Tier não está disponível.
- API do administrador — membro da organização, membro do espaço de trabalho, convite, chave de API, relatório de uso, relatório de custo e pontos finais do relatório de limite de taxa: esses endpoints da API de administração não estão disponíveis no momento. O Workspace CRUD e rename endpoints (`/v1/organizations/workspaces`) estão disponíveis por meio do `aws-external-anthropic` namespace; consulte as ações do IAM. A associação é modelada por meio do AWS IAM e não por meio de listas de membros do espaço de trabalho (consulte a seção anterior). O ciclo de vida da chave de API é gerenciado no Console da AWS em Claude Platform na AWS → Chaves de API. Para dados de uso, custo e limite de taxa, use as visualizações do Claude Console (consulte Monitoramento e registro) ou a API Anthropic Usage and Cost.
- Limites de gastos: Não disponível. Em vez disso, confie nos controles de faturamento da AWS.
- Espaço de trabalho Claude Code e API de análise: o espaço de trabalho Claude Code com limites de taxa automáticos não está disponível. O uso do Claude Code aparece na visualização de uso geral em vez de em uma tela dedicada.
- Autenticação OAuth: não suportada. Use a autenticação SigV4 ou de chave de API.
- OpenAI-compatible Endpoints de API: não disponíveis na plataforma Claude na AWS.
- Workspace-level controles de geografia de inferência: `allowed_inference_geos` e não `default_inference_geo` estão disponíveis. `inference_geo`Em vez disso, defina cada solicitação.

Residência de dados

A plataforma Claude na AWS oferece suporte às seguintes geografias de inferência:

- EUA: A inferência permanece nos data centers dos EUA. Um multiplicador de preços de 1,1x se aplica.
- Global: a inferência pode ser direcionada para qualquer Anthropic-operated data center em todo o mundo. O preço padrão se aplica.

Defina a geografia de inferência por solicitação com o `inference_geo` parâmetro:

> **Note**
>
> O `inference_geo` parâmetro é compatível com Claude Opus 4.6, Claude Sonnet 4.6 e modelos posteriores. Solicitações com `inference_geo` Claude Opus 4.5, Claude Sonnet 4.5 ou Claude Haiku 4.5 retornam um erro 400. Consulte Residência de dados para obter detalhes sobre a disponibilidade do modelo.

Concha

```
curl "https://aws-external-anthropic.us-west-2.api.aws/v1/messages" \
  --aws-sigv4 "aws:amz:us-west-2:aws-external-anthropic" \
  --user "$AWS_ACCESS_KEY_ID:$AWS_SECRET_ACCESS_KEY" \
  -H "x-amz-security-token: $AWS_SESSION_TOKEN" \
  -H "content-type: application/json" \
  -H "anthropic-version: 2023-06-01" \
  -H "anthropic-workspace-id: $ANTHROPIC_AWS_WORKSPACE_ID" \
  -d '{
    "model": "claude-sonnet-4-6",
    "max_tokens": 1024,
    "inference_geo": "us",
    "messages": [
      {"role": "user", "content": "Hello!"}
    ]
  }'
```

Python

```
from anthropic import AnthropicAWS

client = AnthropicAWS(aws_region="us-west-2")
message = client.messages.create(
    model="claude-sonnet-4-6",
    max_tokens=1024,
    inference_geo="us",
    messages=[{"role": "user", "content": "Hello!"}],
)
print(message)
```

TypeScript

```
import AnthropicAws from "@anthropic-ai/aws-sdk";
const client = new AnthropicAws({ awsRegion: "us-west-2" });
const message = await client.messages.create({
  model: "claude-sonnet-4-6",
  max_tokens: 1024,
  inference_geo: "us",
  messages: [{ role: "user", content: "Hello!" }]
});
console.log(message);
```

C#

```
using Anthropic;
using Anthropic.Aws;

var client = new AnthropicAwsClient(new() { AwsRegion = "us-west-2" });

var message = await client.Messages.Create(new()
{
    Model = "claude-sonnet-4-6",
    MaxTokens = 1024,
    InferenceGeo = "us",
    Messages = [new() { Role = "user", Content = "Hello!" }]
});

Console.WriteLine(message);
```

Go

```
client, err := anthropicaws.NewClient(context.Background(),
 anthropicaws.ClientConfig{})
if err != nil {
    panic(err)
}

message, err := client.Messages.New(context.Background(), anthropic.MessageNewParams{
    Model:        anthropic.ModelClaudeSonnet4_6,
    MaxTokens:    1024,
    InferenceGeo: anthropic.String("us"),
    Messages: []anthropic.MessageParam{
        anthropic.NewUserMessage(anthropic.NewTextBlock("Hello!")),
    },
})
if err != nil {
    panic(err)
}

fmt.Println(message)
```

Java

```
import com.anthropic.aws.backends.AwsBackend;
import com.anthropic.client.AnthropicClient;
import com.anthropic.client.okhttp.AnthropicOkHttpClient;
import com.anthropic.models.messages.Message;
import com.anthropic.models.messages.MessageCreateParams;
import com.anthropic.models.messages.Model;
import software.amazon.awssdk.regions.Region;

AnthropicClient client = AnthropicOkHttpClient.builder()
    .backend(AwsBackend.builder().region(Region.of("us-west-2")).build())
    .build();

Message message = client.messages().create(
    MessageCreateParams.builder()
        .model(Model.CLAUDE_SONNET_4_6)
        .maxTokens(1024)
        .inferenceGeo("us")
        .addUserMessage("Hello!")
        .build()
);
```

```
IO.println(message);
```

PHP

```
use Anthropic\Aws\Client;

$client = new Client(awsRegion: 'us-west-2');

$message = $client->messages->create(
    model: 'claude-sonnet-4-6',
    maxTokens: 1024,
    inferenceGeo: 'us',
    messages: [['role' => 'user', 'content' => 'Hello!']],
);

echo $message;
```

Ruby

```
require "anthropic"

client = Anthropic::AWSClient.new(aws_region: "us-west-2")

message = client.messages.create(
  model: "claude-sonnet-4-6",
  max_tokens: 1024,
  inference_geo: "us",
  messages: [{ role: "user", content: "Hello!" }]
)

puts message
```

Se você omitir`inference_geo`, o padrão da solicitação será. `global`

Workspace-level os controles geográficos de inferência (`allowed_inference_geos`e`default_inference_geo`) não estão disponíveis na plataforma Claude na AWS. `inference_geo`Em vez disso, defina cada solicitação.

Espaços de trabalho

As solicitações de inferência e recursos na plataforma Claude na AWS têm como alvo um espaço de trabalho. Você passa o ID do espaço de trabalho no `anthropic-workspace-id` cabeçalho dessas chamadas de API. Os IDs do espaço de trabalho usam o formato marcado `wrkspc_` seguido por um identificador alfanumérico (por exemplo,`wrkspc_01AbCdEf23GhIj`). Consulte Obter seu ID do espaço de trabalho se você ainda não o tiver.

Escopo do espaço de trabalho

Os espaços de trabalho estão vinculados a uma única região da AWS. Um espaço de trabalho criado em só `us-west-2` pode ser acessado por meio do `us-west-2` endpoint. Uso, cotas, custo, arquivos, lotes e habilidades se acumulam por espaço de trabalho, fornecendo detalhamentos por região no Claude Console.

Os espaços de trabalho também servem como o principal recurso do IAM para a plataforma Claude na AWS. Você concede ou nega acesso a espaços de trabalho específicos por meio de políticas do AWS IAM usando o ARN do espaço de trabalho. O segmento de recursos do ARN é o mesmo ID com `wrkspc_` prefixo que você passa no cabeçalho: `anthropic-workspace-id`

```
arn:aws:aws-external-anthropic:{region}:{account-id}:workspace/{workspace-id}
```

Por exemplo: `arn:aws:aws-external-anthropic:us-west-2:123456789012:workspace/wrkspc_01AbCdEf23GhIj`

Consulte as políticas do IAM para ver exemplos de políticas.

Criação de espaços de trabalho

Um espaço de trabalho padrão é provisionado automaticamente no momento da inscrição. Espaços de trabalho adicionais podem ser criados, atualizados, renomeados e arquivados por meio dos `/v1/organizations/workspaces` endpoints, autorizados pelas `aws-external-anthropic` ações correspondentes (, `CreateWorkspaceUpdateWorkspace,ArchiveWorkspace`); consulte as ações do IAM.

Configurando o ID do espaço de trabalho em seu cliente

Cada solicitação do plano de dados deve incluir o ID do espaço de trabalho no `anthropic-workspace-id` cabeçalho. Quando você usa um Claude-on-AWS cliente do Anthropic SDK, há três maneiras de fornecê-lo:

1. Variável de ambiente — definida `ANTHROPIC_AWS_WORKSPACE_ID` e o cliente a lê automaticamente.

   ```
   export ANTHROPIC_AWS_WORKSPACE_ID='wrkspc_01AbCdEf23GhIj'
   ```

2. Argumento do construtor — passe`workspaceId`/`workspace_id`(o nome segue a convenção de linguagem do SDK) ao instanciar o cliente.
3. Per-request substituir — passe o cabeçalho diretamente em uma solicitação individual se precisar abordar vários espaços de trabalho do mesmo cliente.

Se você usa o `Anthropic` cliente base (sem o back-end da AWS), defina o `anthropic-workspace-id` cabeçalho explicitamente em cada solicitação — consulte Como fazer solicitações para exemplos por idioma. Para nomes de SDK-specific argumentos, consulte a documentação dos SDKs do Anthropic Client.

Utilização de tags em espaços de trabalho

As tags do espaço de trabalho são pares de valores-chave anexados a um espaço de trabalho. Eles se integram ao AWS IAM para controle de acesso baseado em atributos e aos relatórios de custo e uso da AWS para alocação de custos (consulte Monitoramento e registro para obter detalhes do CUR). A marcação é GA na plataforma Claude na AWS.

Atualize as tags no espaço de trabalho existente com `TagResource` e `UntagResource` no `aws-external-anthropic` namespace. As mesmas chaves de tag podem orientar as condições do IAM e a alocação de custos sem configuração adicional.

Tag-based controle de acesso

Você pode bloquear `aws-external-anthropic` ações nos valores da tag do espaço de trabalho usando a chave de contexto de `aws:ResourceTag/<key>` condição. A política a seguir permite inferência somente em espaços de trabalho marcados: `Environment=production`

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "aws-external-anthropic:CreateInference",
        "aws-external-anthropic:CreateBatchInference",
        "aws-external-anthropic:CountTokens"
      ],
      "Resource": "arn:aws:aws-external-anthropic:*:*:workspace/*",
      "Condition": {
        "StringEquals": {
          "aws:ResourceTag/Environment": "production"
        }
      }
    }
  ]
}
```

Use `aws:RequestTag/<key>` e `aws:TagKeys` condicione as teclas `TagResource` para aplicar políticas de marcação (por exemplo, exigir que o espaço de trabalho carregue `CostCenter` etiquetas). `Owner` Consulte as políticas do IAM para ver mais exemplos.

Usando o console Claude

A Claude Platform na AWS usa o console Claude padrão em platform.claude.com. Quando você faz login no console da AWS, um indicador de conta gerenciada pela AWS aparece na barra lateral do console Claude. O console se baseia na sua organização Claude Platform na AWS. Ele fornece análises de uso, detalhamento de custos, visibilidade do limite de taxa, visibilidade do espaço de trabalho e páginas para gerenciar arquivos, habilidades do agente, trabalhos em lote, agentes, sessões, ambientes, cofres de credenciais e armazenamentos de memória.

Fazer login

O acesso ao console Claude é federado por meio do AWS IAM. Consulte Configurar sua conta para o fluxo completo de login pela primeira vez. Resumindo:

1. Assuma uma função do IAM com a `aws-external-anthropic:AssumeConsole` permissão. Veja as ações do IAM.
2. Navegue até a página da plataforma Claude na AWS no console da AWS.
3. Escolha Open Claude Console. O AWS Console emite um JWT e redireciona você para o. `platform.claude.com`
4. No primeiro login, você será solicitado a fornecer um endereço de e-mail; insira seu e-mail profissional. A plataforma provisiona seu usuário do Claude Console em tempo hábil.

Duas funções do Claude Console estão disponíveis: administrador e desenvolvedor. A função de administrador concede acesso a todas as páginas e configurações do Claude Console disponíveis para a Claude Platform na AWS. A função Desenvolvedor concede acesso de leitura às informações de uso, custo, limite de taxa e espaço de trabalho. Entre em contato com seu representante de conta da Anthropic para atribuir a função de administrador ou desenvolvedor a um diretor.

Páginas disponíveis

A coluna Via AWS gateway indica se a página lê e grava dados por meio do gateway da AWS (e, portanto, é governada pelas ações do IAM). Páginas marcadas como Não leem metadados de nível organizacional diretamente por meio de APIs antrópicas e ignoram as verificações de ação do IAM.

Página	Available (Disponível)	Por meio do AWS Gateway	Observações
Uso	Sim	Não	Visualize o uso do token por modelo, espaço de trabalho e dimensão. Os dados podem levar alguns minutos para aparecer após uma solicitação.
Custos	Sim	Não	Veja os detalhamentos de custos por modelo e espaço de trabalho. O AWS Cost Explorer mostra o item de linha agregado da CCU.
Limites	Sim	Não	Visualize os limites de taxa (somente leitura).
Espaços de trabalho	Sim	Não	Visualize espaços de trabalho por região (somente leitura).
Arquivos	Sim	Sim	Visualize e gerencie arquivos enviados.
Habilidades	Sim	Sim	Visualize e gerencie as habilidades do agente.
Lotes	Sim	Sim	Visualize e gerencie trabalhos de processam ento em lote.
Atendentes	Sim	Sim	Visualize e gerencie as definições do agente.
Sessões	Sim	Sim	Veja as sessões do agente e o histórico de eventos.
Ambientes do	Sim	Sim	Visualize e gerencie as configurações de contêineres na nuvem para sessões.
Cofres de credenciais	Sim	Sim	Visualize e gerencie cofres de credenciais para autenticação de sessão.
Armazenam entos de memória	Sim	Sim	Visualize e gerencie a memória persistente do agente.

Página	Available (Disponível)	Por meio do AWS Gateway	Observações
Chaves de API	Não	N/A	Gerencie chaves de API no Console da AWS (Claude Platform na AWS → Chaves de API). Consulte Autenticação.
Membros	Não	N/A	Não aplicável. O AWS IAM gerencia o acesso.
Faturamento	Não	N/A	Não aplicável. O AWS Marketplace gerencia o faturamento e o faturamento. Veja os detalhamentos dos custos na página Custo.
Código Claude	Não	N/A	Veja o uso do Claude Code na página Uso.

Mudando de organização

O console Claude não oferece suporte à mudança de organizações para a plataforma Claude na AWS. Para acessar uma organização diferente, saia e reautentique-se por meio do AWS Console usando a função IAM da conta da AWS dessa organização.

Limites de tarifas e cotas

A Claude Platform na AWS atribui limites de taxa de nível 1 quando você se inscreve. A Anthropic gerencia os limites de taxa diretamente, não por meio dos sistemas de cotas da AWS.

Limites padrão

A Claude Platform na AWS usa a programação de níveis padrão da Anthropic, idêntica à API Claude primária. Os limites de nível 1 se aplicam por espaço de trabalho. Os limites são agrupados por família de modelos. Por exemplo, todos os modelos Opus compartilham um limite combinado, todos os modelos Sonnet compartilham outro e todos os modelos Haiku compartilham um terceiro.

Para valores atuais de nível 1 (RPM, ITPM, OTPM) e limites de nível superior, consulte Limites de taxa no site de documentação da Anthropic. A página Anthropic é a fonte da verdade e é atualizada quando os limites mudam.

Cabeçalhos de limite de taxa

Cada resposta inclui cabeçalhos que informam o status atual do seu limite de taxa. Cabeçalhos principais:

- `anthropic-ratelimit-requests-limit`— Máximo de solicitações por minuto
- `anthropic-ratelimit-requests-remaining`— Solicitações restantes na janela atual
- `anthropic-ratelimit-requests-reset`— Hora em que o limite da solicitação é redefinido (RFC 3339)
- `anthropic-ratelimit-tokens-limit`— Máximo de tokens combinados (entrada + saída) por minuto
- `anthropic-ratelimit-tokens-remaining`— Tokens combinados restantes na janela atual
- `anthropic-ratelimit-tokens-reset`— Hora em que o limite combinado de tokens é redefinido (RFC 3339)
- `anthropic-ratelimit-input-tokens-limit`/`-remaining`/`-reset`— Input-token-specific cabeçalhos
- `anthropic-ratelimit-output-tokens-limit`/`-remaining`/`-reset`— Output-token-specific cabeçalhos

- `retry-after`— Em uma resposta 429, o número de segundos de espera antes de tentar novamente

Consulte os cabeçalhos de resposta no site da documentação da Anthropic para ver o conjunto completo.

Solicitando limites mais altos

Ao contrário da API Claude original, o avanço automático de nível não se aplica à plataforma Claude na AWS. Para solicitar limites mais altos, entre em contato com seu representante de conta da Anthropic com seu ID do espaço de trabalho e a taxa de transferência desejada. Para ver os limites de nível e outros detalhes, consulte Limites de tarifas no site de documentação da Anthropic.

Erros de limite de taxa

Quando você excede um limite de taxa, a API retorna HTTP 429 com um `rate_limit_error` tipo. Implemente um recuo exponencial com instabilidade em sua lógica de repetição. O `retry-after` cabeçalho indica quantos segundos esperar antes de tentar novamente.

Faturamento

A Claude Platform na AWS usa o AWS Marketplace como um back-end de cobrança para medição baseada em consumo. A Anthropic mede o uso de hora em hora por meio do AWS Marketplace, e a AWS emite faturas mensais sobre sua fatura existente da AWS.

O faturamento é somente em atraso (você paga pelo que usou) e antes de impostos (a AWS aplica as configurações fiscais da sua conta).

O uso é denominado em Unidades de Consumo Claude (CCUs) a 0,01 USD por CCU. O preço da CCU é fixo e nunca é descontado. A Anthropic avalia seu uso de token em USD de acordo com as tarifas padrão por modelo e por recurso, aplica qualquer desconto negociado e, em seguida, converte o resultado em CCUs a 0,01 USD por CCU. Os descontos resultam em menos CCUs medidas, não em um preço de CCU mais baixo. As CCUs não são créditos pré-pagos; não há saldo ou compromisso da CCU. Consulte Preços para ver a definição da CCU e as taxas de token por modelo.

Monitorar e registrar

A AWS CloudTrail pode capturar todas as solicitações para a plataforma Claude na AWS. As operações do espaço de trabalho e do cofre são registradas como eventos de gerenciamento por padrão. Todas as outras operações (inferência, lote, arquivo, habilidade, modelo, perfil de usuário e agentes gerenciados por Claude, exceto cofres) são eventos de dados. Registrá-los requer configuração explícita e acarreta custos adicionais CloudTrail . Consulte as ações do IAM para obter a classificação completa do tipo de evento e a CloudTrail documentação da AWS para obter detalhes de configuração.

Ao configurar o registro de eventos de CloudTrail dados, selecione o tipo `AWS::AWSExternalAnthropic::Workspace` de recurso. Esse é o tipo de CloudTrail recurso da plataforma Claude nos espaços de trabalho da AWS e é necessário para capturar eventos do plano de dados (inferência, lote, arquivo e outras operações por espaço de trabalho). Defina o escopo dos seletores de eventos de dados para ARNs de espaços de trabalho específicos se você quiser registrar atividades em um subconjunto de espaços de trabalho em vez de em toda a conta.

Solicitar IDs

Cada resposta inclui duas IDs de solicitação nos cabeçalhos de resposta:

- ID de solicitação da AWS (**`x-amzn-requestid`**): o ID principal, indexado em CloudTrail. Use isso ao investigar solicitações por meio das ferramentas da AWS ou ao entrar em contato com o suporte da AWS.
- ID de solicitação antrópica (**`request-id`**): a ID secundária. Use isso ao entrar em contato com o suporte da Anthropic.

Python

```
from anthropic import AnthropicAWS

client = AnthropicAWS(aws_region="us-west-2")

response = client.messages.with_raw_response.create(
    model="claude-sonnet-4-6",
    max_tokens=1024,
    messages=[{"role": "user", "content": "Hello!"}],
)
```

```
print(response.headers.get("x-amzn-requestid"))  # AWS request ID
print(response.headers.get("request-id"))  # Anthropic request ID

message = response.parse()
print(message.content)
```

TypeScript

```
import AnthropicAws from "@anthropic-ai/aws-sdk";

const client = new AnthropicAws({ awsRegion: "us-west-2" });

const { data: message, response } = await client.messages
  .create({
    model: "claude-sonnet-4-6",
    max_tokens: 1024,
    messages: [{ role: "user", content: "Hello!" }]
  })
  .withResponse();

console.log(response.headers.get("x-amzn-requestid")); // AWS request ID
console.log(response.headers.get("request-id")); // Anthropic request ID
console.log(message.content);
```

C#

```
using Anthropic;
using Anthropic.Aws;

var client = new AnthropicAwsClient(new() { AwsRegion = "us-west-2" });

var response = await client.WithRawResponse.Messages.Create(new()
{
    Model = "claude-sonnet-4-6",
    MaxTokens = 1024,
    Messages = [new() { Role = "user", Content = "Hello!" }]
});

Console.WriteLine(response.Headers.GetValues("x-amzn-requestid").First()); // AWS
 request ID
Console.WriteLine(response.Headers.GetValues("request-id").First()); // Anthropic
 request ID
```

```
Console.WriteLine(response.Value.Content);
```

Go

```
client, err := anthropicaws.NewClient(context.Background(),
 anthropicaws.ClientConfig{})
if err != nil {
    panic(err)
}

var response *http.Response
message, err := client.Messages.New(
    context.Background(),
    anthropic.MessageNewParams{
        Model:     anthropic.ModelClaudeSonnet4_6,
        MaxTokens: 1024,
        Messages: []anthropic.MessageParam{
            anthropic.NewUserMessage(anthropic.NewTextBlock("Hello!")),
        },
    },
    option.WithResponseInto(&response),
)
if err != nil {
    panic(err)
}

fmt.Println(response.Header.Get("x-amzn-requestid")) // AWS request ID
fmt.Println(response.Header.Get("request-id"))       // Anthropic request ID
fmt.Println(message.Content[0].Text)
```

Java

```
import com.anthropic.aws.backends.AwsBackend;
import com.anthropic.client.AnthropicClient;
import com.anthropic.client.okhttp.AnthropicOkHttpClient;
import com.anthropic.core.http.HttpResponseFor;
import com.anthropic.models.messages.Message;
import com.anthropic.models.messages.MessageCreateParams;
import com.anthropic.models.messages.Model;

AnthropicClient client = AnthropicOkHttpClient.builder()
    .backend(AwsBackend.fromEnv())
    .build();
```

```
HttpResponseFor<Message> response = client.messages().withRawResponse().create(
    MessageCreateParams.builder()
        .model(Model.CLAUDE_SONNET_4_6)
        .maxTokens(1024)
        .addUserMessage("Hello!")
        .build()
);

IO.println(response.headers().values("x-amzn-requestid")); // AWS request ID
IO.println(response.requestId()); // Anthropic request ID
IO.println(response.parse().content());
```

PHP

```
use Anthropic\Aws\Client;

$client = new Client();

$response = $client->messages->raw->create(
    model: 'claude-sonnet-4-6',
    maxTokens: 1024,
    messages: [['role' => 'user', 'content' => 'Hello!']],
);

echo $response->getHeaderLine('x-amzn-requestid') . "\n"; // AWS request ID
echo $response->getHeaderLine('request-id') . "\n"; // Anthropic request ID
echo $response->parse();
```

Ruby

Atualmente, o SDK do Ruby não expõe um acessador de resposta bruta, portanto, os IDs de solicitação não podem ser lidos nos cabeçalhos de resposta em Ruby. Se você precisar de registro de ID de solicitação, use qualquer uma das outras linguagens acima ou capture as IDs na camada de proxy HTTP.

A Anthropic recomenda registrar sua atividade em pelo menos 30 dias contínuos para entender os padrões de uso e investigar possíveis problemas.

> **Note**
>
> CloudTrail A AWS está configurada na sua conta da AWS. A ativação do registro em log não fornece à AWS ou à Anthropic acesso ao seu conteúdo além do necessário para a operação de faturamento e serviços.

Métricas e painéis de uso

A Claude Platform na AWS não publica métricas de uso por espaço de trabalho na Amazon. CloudWatch Em vez disso, a contagem de tokens, o volume de solicitações e o uso por modelo estão disponíveis nas superfícies de uso da Anthropic:

- Visualizações de uso do Claude Console — quando um principal se federa no Claude Console com `aws-external-anthropic:AssumeConsole` (consulte as políticas do IAM), os painéis de uso mostram o volume de solicitações, a contagem de tokens e os detalhamentos por modelo para espaços de trabalho acessíveis. Account-wide as visualizações de uso exigem o recurso do console administrativo.
- API de uso e custo antrópicos — para acesso programático aos dados de uso e custo, incluindo agregação por espaço de trabalho e por modelo, consulte API de uso e custo na documentação da Anthropic.

Para monitoramento operacional (taxas de erro, latência, cabeçalhos de limite de taxa), use os cabeçalhos de resposta retornados em cada solicitação (consulte Limites de taxa e cotas) junto com os IDs de solicitação da AWS capturados. CloudTrail

Alocação e monitoramento de custos

As cobranças da Plataforma Claude na AWS aparecem em sua fatura da AWS no serviço Claude Platform na AWS, com dimensões de uso por modelo. Use o Relatório de Custos e Uso da AWS (CUR) combinado com tags de espaço de trabalho para alocação de custos refinada:

1. Marque seus espaços de trabalho com as dimensões de alocação que lhe interessam (equipe, aplicativo, ambiente, centro de custos). Consulte Espaços de trabalho para obter detalhes sobre a marcação.
2. No console de faturamento da AWS, ative cada chave de tag como uma tag de alocação de custos. Após a ativação, o uso incorrido em ou após a data de ativação é dividido por tag no CUR.

3. No CUR ou no AWS Cost Explorer, agrupe pela `resourceTags/user:<tag-key>` coluna para detalhar os gastos por tag do espaço de trabalho.

As tags do espaço de trabalho fluem para os itens da linha CUR para todo o uso da plataforma Claude na AWS. As mesmas chaves de tag orientam tanto o controle de IAM-based acesso quanto a alocação de custos. Consulte as tags de alocação de custos na documentação do AWS Billing para ver as etapas de configuração e os atrasos na ativação.

Migração do Amazon Bedrock

Se você usa atualmente o Claude no Bedrock, a migração para a plataforma Claude na AWS exige mudanças em toda a sua integração. A assinatura SigV4 continua suportada, mas o contexto da assinatura, o URL base, o formato da API, os IDs do modelo, o cliente e o pacote do SDK, o formato de streaming, os cabeçalhos da solicitação e a disponibilidade da região mudam. A tabela a seguir resume as diferenças.

O que muda

	Amazon Bedrock	Plataforma Claude na AWS
URL base	`bedrock-runtime.{region}.amazonaws.com`	`aws-external-anthropic.{region}.api.aws`
Formato da API	Bedrock Converse/ InvokeModel	API de mensagens antrópicas () `/v1/messages`
IDs do modelo	`anthropic.claude-opus-4-7-v1`	`claude-opus-4-7`
Cliente SDK	`AnthropicBedrock` /SDK Bedrock	Platform-specific cliente (`AnthropicAWS` ,`AnthropicAws` ,`AnthropicAwsClient` , etc.), em versão beta
Pacote SDK	`anthropic[bedrock]` ,`@anthropic-ai/bedrock-sdk`, etc.	`anthropic[aws]` `@anthropic-ai/aws-sdk` , etc. (consulte Instalar um SDK)
Nome do serviço SigV4	`bedrock`	`aws-external-anthropic`
Formato de streaming	AWS EventStream	SSE (o mesmo que Claude API)

	Amazon Bedrock	Plataforma Claude na AWS
cabeçalho do espaço de trabalho	Não aplicável	`anthropic-workspace-id` obrigatório
Disponibilidade de regiões	Várias regiões comerciais	Todas as regiões comerciais da AWS (as regiões opt-in exigem a aceitação da conta)

O que você ganha

- Normalmente, acesso no mesmo dia a novos modelos e recursos, sem uma etapa separada de integração com parceiros
- Habilidades do agente para geração de documentos (ExcelPowerPoint, Word, PDF)
- Execução de código na sandbox gerenciada da Anthropic
- Recursos beta por meio do `anthropic-beta` cabeçalho (consulte Recursos não disponíveis no momento)
- Claude Console para visibilidade de cotas e análise de uso
- Suporte antrópico direto
- Autenticação de chave de API como alternativa ao SigV4 (consulte Autenticação)

O que permanece o mesmo

- Autenticação AWS IAM (SigV4)
- Faturamento por meio de sua conta da AWS
- Aposentadoria do compromisso da AWS

Armadilhas da migração

> **Warning**
>
> Ative primeiro a federação de identidade da web de saída. Se sua conta da AWS não usou anteriormente a plataforma Claude na AWS, você deve habilitar a federação de identidade web de saída uma vez por conta antes de fazer solicitações. Sem essa etapa, todas as solicitações falham com um erro de federação. Esta etapa não é necessária para o Bedrock.

> **Warning**
>
> A retenção zero de dados (ZDR) é opcional na plataforma Claude na AWS. No Bedrock, a AWS é a processadora de dados e a Anthropic não retém entradas ou saídas de inferência; o programa ZDR da Anthropic não se aplica lá. Na plataforma Claude na AWS, a Anthropic é a processadora de dados, e a ZDR segue o modelo primário da API Claude: ela está disponível mediante solicitação por meio de seu representante de conta da Anthropic. Confirme a inscrição no ZDR antes de migrar cargas de trabalho de produção que dependem das garantias de retenção de dados.

Considerações comerciais

- Termos de serviço: Os Termos de Serviço da AWS regem a Plataforma Claude na AWS. Os Termos de Serviço Comerciais, o Adendo de Processamento de Dados e a Política de Uso da Anthropic também se aplicam.
- Descontos e ofertas privadas: descontos negociados e ofertas privadas do AWS Marketplace não são transferidos automaticamente entre as plataformas Bedrock e Claude na AWS. Trabalhe com seu representante de conta da Anthropic para configurar termos comerciais para a plataforma Claude na AWS.

Políticas do IAM

A plataforma Claude na AWS se integra ao AWS IAM para controle de acesso. Você concede ou nega acesso a ações específicas da API em espaços de trabalho específicos usando a sintaxe padrão da política do IAM.

O nome do serviço SigV4 e o namespace da ação do IAM são. `aws-external-anthropic` As ações seguem o padrão `aws-external-anthropic:<Action>` (por exemplo,`aws-external-anthropic:CreateInference`).

Exemplo: negar inferência em lote

A política a seguir permite inferência em tempo real enquanto bloqueia o processamento em lote, um requisito comum para ZDR-sensitive cargas de trabalho:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "aws-external-anthropic:CreateInference",
        "aws-external-anthropic:CountTokens",
        "aws-external-anthropic:GetModel",
        "aws-external-anthropic:ListModels",
        "aws-external-anthropic:GetWorkspace",
        "aws-external-anthropic:ListWorkspaces"
      ],
      "Resource": "arn:aws:aws-external-anthropic:*:*:workspace/*"
    },
    {
      "Effect": "Deny",
      "Action": [
        "aws-external-anthropic:CreateBatchInference",
        "aws-external-anthropic:GetBatchInference",
        "aws-external-anthropic:ListBatchInferences"
      ],
      "Resource": "*"
    }
  ]
```

```
}
```

A `GetBatchInference` ação autoriza a rota de metadados em lote e a rota de resultados em lote. Além `ListBatchInferences` disso, negá-lo bloqueia as leituras e a enumeração em lote.

A `Allow` declaração enumera `List*` ações específicas em vez de `Get*` usar curingas. Os curingas concederiam `GetFile` (que baixam os bytes do arquivo) e outras leituras que você talvez não pretenda; `Deny` substituem de `Allow` qualquer maneira, mas a forma explícita é o padrão mais seguro para modelar.

Exemplo: inferência síncrona em um único espaço de trabalho

Concede as permissões mínimas para um diretor do IAM que executa inferências em um espaço de trabalho de produção:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "aws-external-anthropic:CreateInference",
        "aws-external-anthropic:CountTokens",
        "aws-external-anthropic:Get*",
        "aws-external-anthropic:List*"
      ],
      "Resource": "arn:aws:aws-external-anthropic:us-west-2:123456789012:workspace/
wrkspc_01AbCdEf23GhIj"
    }
  ]
}
```

> **Note**
>
> O `List*` caractere curinga nessa política também corresponde`ListWorkspaces`, que tem como escopo a conta. A restrição ARN do espaço de trabalho a filtra silenciosamente, portanto, essa política não autoriza a listagem de espaços de trabalho. Se sua conta de serviço precisar enumerar espaços de trabalho, adicione uma declaração separada `Allow` para with. `ListWorkspaces` `Resource: "*"`

Essa política pressupõe a autenticação AWS SigV4. Se o principal se autenticar com uma chave de API, conceda também `aws-external-anthropic:CallWithBearerToken` (consulte Autenticação).

Exemplo: isolamento do espaço de trabalho por cliente

Restringe uma função a um único espaço de trabalho:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "aws-external-anthropic:*",
      "Resource": "arn:aws:aws-external-anthropic:us-west-2:123456789012:workspace/
wrkspc_01AbCdEf23GhIj"
    },
    {
      "Effect": "Allow",
      "Action": [
        "aws-external-anthropic:CallWithBearerToken",
        "aws-external-anthropic:AssumeConsole"
      ],
      "Resource": "*"
    }
  ]
}
```

Note

O `aws-external-anthropic:*` curinga na primeira declaração inclui ações com escopo de conta (`CreateWorkspace`,`ListWorkspaces`) que a restrição ARN do espaço de trabalho filtra silenciosamente. Isso é consistente com a intenção de “isolamento” — a função não pode criar ou enumerar espaços de trabalho — mas a política contém permissões que não têm efeito. Consulte Automação de provisionamento para ver o padrão de escopo da conta.

A segunda declaração concede `CallWithBearerToken` e `AssumeConsole` sobre todos os recursos porque ambas são ações sem rota que não se vinculam a um ARN do espaço de

trabalho. Omita a segunda declaração se a função usar somente SigV4 e nunca se federar no Claude Console.

Exemplo: bloqueio de recursos para um espaço de trabalho ZDR-sensitive

Bloqueia o processamento em lote e o upload de arquivos em um espaço de trabalho específico, deixando a inferência síncrona disponível. Útil quando um espaço de trabalho manipula dados de retenção zero de dados (ZDR) que não devem persistir no lado do servidor. Anexe essa política junto com uma política de Permitir, como `AnthropicLimitedAccess` o exemplo de espaço de trabalho único acima; sozinha, uma Deny-only política não concede permissões:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Deny",
      "Action": [
        "aws-external-anthropic:CreateBatchInference",
        "aws-external-anthropic:CreateFile"
      ],
      "Resource": "arn:aws:aws-external-anthropic:us-west-2:123456789012:workspace/
wrkspc_01AbCdEf23GhIj"
    }
  ]
}
```

> **Note**
>
> Essa negação bloqueia somente a criação. Outras ações de arquivo e lote não são negadas, a menos que você também as liste. Para um bloqueio completo em que o espaço de trabalho nunca deve conter arquivos ou lotes, negue também`aws-external-anthropic:GetFile`,,`aws-external-anthropic:ListFiles`,`aws-external-anthropic:DeleteFile`, `aws-external-anthropic:GetBatchInferenceaws-external-anthropic:ListBatchInferences`, e. `aws-external-anthropic:CancelBatchInference` `aws-external-anthropic:DeleteBatchInference`

Exemplo: automação de provisionamento

Concede a uma CI/CD função as ações necessárias para criar e gerenciar espaços de trabalho, sem nenhuma permissão de inferência:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "aws-external-anthropic:CreateWorkspace",
        "aws-external-anthropic:GetWorkspace",
        "aws-external-anthropic:ListWorkspaces",
        "aws-external-anthropic:UpdateWorkspace",
        "aws-external-anthropic:ArchiveWorkspace"
      ],
      "Resource": "*"
    }
  ]
}
```

`CreateWorkspacee` `ListWorkspaces` são operações com escopo de conta. Especificar um ARN do espaço de trabalho nessas ações não tem efeito; use. `Resource: "*"`

Políticas gerenciadas

A AWS fornece políticas gerenciadas para padrões de acesso comuns:

- **`AnthropicFullAccess`**: Subsídios `aws-external-anthropic:*` em todos os recursos.
- **`AnthropicReadOnlyAccess`**: `List*` Subsídios `Get*` e `CallWithBearerToken` sobre todos os recursos.
- **`AnthropicInferenceAccess`**: concede as ReadOnly ações mais as ações de inferência (`CreateInference`,,`CreateBatchInference`, `CancelBatchInferenceDeleteBatchInference`,`CountTokens`) em todos os recursos.
- **`AnthropicLimitedAccess`**: concede as `AnthropicInferenceAccess` ações mais todas as ações dos Agentes Gerenciados Claude (agentes, sessões, ambientes, cofres, armazenamentos de memória) em todos os recursos.

- **`AnthropicSelfHostedEnvironmentAccess`**: concede as permissões que um trabalhador de sandbox auto-hospedado precisa para pesquisar e processar itens de trabalho do ambiente, ler o ambiente, a sessão e os recursos de habilidades associados e atualizar o estado da sessão. Vincule essa política à função do IAM que seu trabalhador de ambiente auto-hospedado assume. `AnthropicSelfHostedEnvironmentAccess`é o padrão de função única recomendado; se você executar o questionário de trabalho e o sandbox por sessão em diretores do IAM separados, poderá dividir essas permissões em duas políticas personalizadas mais restritas para obter o menor privilégio.

`AnthropicInferenceAccess`é a política gerenciada mais restrita suficiente para executar inferências. Por meio dos `List*` curingas `Get*` e, ele concede acesso de leitura a todos os recursos da API no namespace, incluindo o download do conteúdo do arquivo `GetFile` e o conteúdo da memória. `GetMemoryStore` Ela não concede a criação ou exclusão de arquivos ou habilidades, o gerenciamento do perfil do usuário, a mutação do espaço de trabalho ou a federação do console.

> **Note**
>
> `AnthropicReadOnlyAccess`,`AnthropicInferenceAccess`, e `AnthropicLimitedAccess` não concedam`AssumeConsole`. Os diretores que precisam se associar ao Claude Console precisam de uma concessão separada para`aws-external-anthropic:AssumeConsole`, por meio de uma política personalizada `AnthropicFullAccess` ou por meio de uma política personalizada. Consulte Federando no console Claude.

> **Note**
>
> `CreateInference`e `CreateBatchInference` são ações separadas. Negar um não bloqueia o outro. Se você pretende evitar todas as chamadas de modelo, negue ambas.

Federando para o Claude Console

`aws-external-anthropic:AssumeConsole`permite que um diretor do IAM seja federado no Anthropic-operated Claude Console. O acesso ao console ainda é controlado pelo IAM para a maioria das operações, mas um subconjunto de operações administrativas — principalmente

visualizações de uso que não têm a API IAM correspondente — é limitado pela capacidade com a qual o principal está federado.

Existem dois recursos:

- **`developer`**— permite as operações que um desenvolvedor da Claude Platform na AWS precisa para o trabalho diário: executar inferências a partir do console, ler dados do espaço de trabalho, visualizar o uso pessoal.
- **`admin`**— além disso, permite operações de console somente para administradores, incluindo visualizações de uso em toda a conta e configurações administrativas que não são exibidas por meio de ações do IAM.

Controle qual capacidade um diretor pode solicitar com a chave de `aws-external-anthropic:Capability` condição na `AssumeConsole` ação:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "aws-external-anthropic:AssumeConsole",
      "Resource": "*",
      "Condition": {
        "StringEquals": {
          "aws-external-anthropic:Capability": "admin"
        }
      }
    }
  ]
}
```

`AnthropicFullAccess`subsídios `AssumeConsole` sem restrição de capacidade. Para qualquer concessão mais restrita — acesso somente para desenvolvedores ao console ou acesso somente para administradores — anexe uma política personalizada que atenda à condição, conforme mostrado acima. `AssumeConsole` `aws-external-anthropic:Capability`

Chamando com fichas de portador

`aws-external-anthropic:CallWithBearerToken`autoriza o caminho da SigV4-free solicitação usado quando uma chave de API é apresentada como um token portador. Qualquer principal que se autentique com uma chave de API precisa dessa ação no espaço de trabalho de destino. Isso se aplica se você usa `ANTHROPIC_AWS_API_KEY` ou define o `Authorization: Bearer` cabeçalho diretamente.

Isso é necessário para chamadores de chave de API, além das ações de inferência (`CreateInference`,`CreateBatchInference`, etc.). Sem isso`CallWithBearerToken`, as solicitações de chave de API são rejeitadas antes de chegarem à verificação de autorização de inferência. Os chamadores do SigV4 não precisam dessa ação.

`AnthropicReadOnlyAccess`,`AnthropicInferenceAccess`,`AnthropicLimitedAccess`, e `AnthropicFullAccess` tudo incluído`CallWithBearerToken`. Se você escrever uma política personalizada para acesso à chave de API, adicione-a explicitamente.

Federação de identidade da web de saída (necessária para acesso ao console)

O Claude Console é executado na infraestrutura antrópica, não na AWS. Quando um diretor do IAM liga`AssumeConsole`, o AWS STS emite um token de identidade da web com escopo definido para o público antrópico; o console do Anthropic então aceita esse token e estabelece a sessão federada. Para que isso funcione, a conta da AWS deve permitir a federação externa de identidade da web para o `aws-external-anthropic` público.

Os diretores que ligam `AssumeConsole` precisam das seguintes permissões STS, além da `aws-external-anthropic:AssumeConsole` ação:

- **`sts:GetWebIdentityToken`**— permite a emissão do token de identidade da web que o console Anthropic consome.
- **`sts:TagGetWebIdentityToken`**— permite anexar tags de sessão ao token de identidade da web. A Claude Platform na AWS usa essas tags para transmitir a capacidade e o contexto do espaço de trabalho do diretor ao console.

Inclua tanto na política de confiança de qualquer função que os usuários do console assumam quanto na inline/managed política anexada às identidades dos usuários que ligam `AssumeConsole`

diretamente. `AnthropicFullAccess`inclui as duas ações do STS. Ambientes que negam `sts:*` no nível do SCP ou do limite de permissão devem permitir explicitamente essas duas ações para que a federação do console seja bem-sucedida.

Ações do IAM para a plataforma Claude na AWS

Referência de ação do IAM para controlar o acesso à plataforma Claude na AWS por meio de políticas da AWS.

A Claude Platform na AWS usa o AWS IAM para controle de acesso. Cada rota de API é mapeada para uma ação do IAM no `aws-external-anthropic` namespace. Esta página lista todas as ações, as rotas autorizadas por cada ação e as políticas gerenciadas disponíveis para padrões de acesso comuns. Para configuração e autenticação da plataforma, consulte O que é a plataforma Claude na AWS? .

Detalhes do serviço

Prefixo do serviço IAM	`aws-external-anthropic`
Resource types	`workspace`
ARN do espaço de trabalho	`arn:aws:aws-external-anthropic:{region}:{account-id}:workspace/{workspace-id}`

A região do ARN é sempre preenchida e corresponde à região à qual o espaço de trabalho está vinculado. O segmento do recurso é o ID do espaço de trabalho marcado (`wrkspc_…`), o mesmo valor que você passa no `anthropic-workspace-id` cabeçalho.

Ações

O serviço define 65 ações em 15 grupos. As ações seguem a `VerbNoun` convenção da AWS e usam a disciplina verbal para que os `Get*` `List*` curingas produzam um limite claro somente para leitura.

Inferência

Ação	Rotas autorizadas
`CreateInference`	`POST /v1/messages`

Ação	Rotas autorizadas
CountTokens	POST /v1/messages/count_tokens

Processamento em lotes

Ação	Rotas autorizadas
CreateBatchInference	POST /v1/messages/batches
GetBatchInference	GET /v1/messages/batches/{id} GET /v1/messages/batches/{id}/results
ListBatchInferences	GET /v1/messages/batches
CancelBatchInference	POST /v1/messages/batches/{id}/cancel
DeleteBatchInference	DELETE /v1/messages/batches/{id}

Modelos da

Ação	Rotas autorizadas
GetModel	GET /v1/models/{id}
ListModels	GET /v1/models

Arquivos

Ação	Rotas autorizadas
CreateFile	POST /v1/files
GetFile	GET /v1/files/{id} GET /v1/files/{id}/content
ListFiles	GET /v1/files

Ação	Rotas autorizadas
`DeleteFile`	`DELETE /v1/files/{id}`

> **Note**
>
> `GetFile`autoriza o download de metadados e conteúdo. Um diretor com acesso somente para leitura pode baixar bytes de arquivos, não apenas listar arquivos.

Habilidades

Ação	Rotas autorizadas
`CreateSkill`	`POST /v1/skills`
`GetSkill`	`GET /v1/skills/{id}` `GET /v1/skills/{id}/versions` `GET /v1/skills/{id}/versions/{version}` `GET /v1/skills/{id}/versions/{version}/content`
`ListSkills`	`GET /v1/skills`
`UpdateSkill`	`POST /v1/skills/{id}/versions` `DELETE /v1/skills/{id}/versions/{version}`
`DeleteSkill`	`DELETE /v1/skills/{id}`

> **Note**
>
> A exclusão de uma versão de habilidade individual é mapeada para`UpdateSkill`, não`DeleteSkill`. Uma política que nega `aws-external-anthropic:Delete*` ainda permite a exclusão da versão. `UpdateSkill`Negue `CreateSkill` também se precisar evitar qualquer mutação de habilidade.

Perfis de usuário

Ação	Rotas autorizadas
`CreateUserProfile`	`POST /v1/user_profiles`
`GetUserProfile`	`GET /v1/user_profiles/{id}`
`ListUserProfiles`	`GET /v1/user_profiles`
`UpdateUserProfile`	`POST /v1/user_profiles/{id}`

> **Warning**
>
> A correspondência de ações do IAM não diferencia maiúsculas de minúsculas. O caractere curinga `aws-external-anthropic:*File` corresponde a `CreateFileGetFile`, e`DeleteFile`, mas não corresponde `ListFiles` (o que termina em “arquivos”, não em “arquivo”). Também coincide demais com`CreateUserProfile`,`GetUserProfile`, e `UpdateUserProfile` porque “Perfil” termina em “arquivo”. Se você pretende conceder ou negar somente ações da API Files, enumere-as explicitamente (`CreateFile`,, `GetFileListFiles`,`DeleteFile`) em vez de usar um padrão de sufixo. `*File`

Agentes

Definições de agente para Claude Managed Agents.

Ação	Rotas autorizadas
`CreateAgent`	Agente cria rotas
`GetAgent`	Rotas de leitura do agente
`ListAgents`	Rotas da lista de agentes
`UpdateAgent`	Rotas de atualização do agente
`ArchiveAgent`	rotas de arquivamento do agente

Sessões

Sessões do agente e histórico de eventos.

Ação	Rotas autorizadas
`CreateSession`	Sessão: criar rotas
`GetSession`	Rotas de leitura da sessão
`ListSessions`	Rotas da lista de sessões
`UpdateSession`	Rotas de atualização da sessão
`ArchiveSession`	Rotas de arquivamento de
`DeleteSession`	Rotas de exclusão de sessão

Ambientes do

Configurações de contêiner de nuvem para sessões.

Ação	Rotas autorizadas
`CreateEnvironment`	Ambiente: crie rotas
`GetEnvironment`	Rotas de leitura do ambiente
`ListEnvironments`	Rotas da lista de ambientes
`UpdateEnvironment`	Rotas de atualização do ambiente
`ArchiveEnvironment`	Rotas de arquivamento ambiental
`DeleteEnvironment`	Ambiente: excluir rotas
`ProcessEnvironment Work`	Self-hosted rotas de trabalhadores ambientais

Cofres

Cofres de credenciais para autenticação de sessão.

Ação	Rotas autorizadas
`CreateVault`	Vault cria rotas
`GetVault`	Rotas de leitura do Vault
`ListVaults`	Rotas da lista do Vault
`UpdateVault`	Rotas de atualização do Vault
`ArchiveVault`	Rotas de arquivamento do Vault
`DeleteVault`	Rotas de exclusão do Vault

> **Note**
>
> As operações do cofre são classificadas como eventos CloudTrail de gerenciamento (em vez de eventos de dados) porque os cofres contêm credenciais e se beneficiam do registro de auditoria padrão. Outras operações do Claude Managed Agents (agentes, sessões, ambientes, armazenamentos de memória) são eventos de dados.

Armazenamentos de memória

Memória persistente do agente.

Ação	Rotas autorizadas
`CreateMemoryStore`	Armazenamento de memória: crie rotas
`GetMemoryStore`	Rotas de leitura do armazenamento de memória
`ListMemoryStores`	Rotas da lista de armazenamento de memória
`UpdateMemoryStore`	Rotas de atualização do armazenamento de memória

Ação	Rotas autorizadas
`ArchiveMemoryStore`	Rotas de arquivamento da memória
`DeleteMemoryStore`	Rotas de exclusão do armazenamento de memória

> **Note**
>
> `GetMemoryStore`lê o conteúdo da memória. O `Get*` caractere curinga em uma política gerenciada ou personalizada, portanto, concede acesso de leitura à memória. Políticas de escopo explícitas se o conteúdo da memória precisar ser restrito.

Webhooks

Notificações de eventos de ciclo de vida para sessões.

Ação	Rotas autorizadas
`ListWebhooks`	`GET /v1/webhooks`
`GetWebhook`	`GET /v1/webhooks/{webhook_endpoint_id}`
`CreateWebhook`	`POST /v1/webhooks`
`UpdateWebhook`	`POST /v1/webhooks/{webhook_endpoint_id}`
`DeleteWebhook`	`DELETE /v1/webhooks/{webhook_endpoint_id}`
`RotateWebhookSecret`	`POST /v1/webhooks/{webhook_endpoint_id}/regenerate_signing_secret`

Espaços de trabalho

Ação	Rotas autorizadas
`CreateWorkspace`	`POST /v1/organizations/workspaces`

Ação	Rotas autorizadas
`GetWorkspace`	`GET /v1/organizations/workspaces/{id}`
`ListWorkspaces`	`GET /v1/organizations/workspaces`
`UpdateWorkspace`	`POST /v1/organizations/workspaces/{id}`
`ArchiveWorkspace`	`POST /v1/organizations/workspaces/{id}/archive`

Um espaço de trabalho padrão é provisionado no momento da inscrição; consulte Espaços de trabalho.

Autenticação

Ação	Rotas autorizadas
`CallWithBearerToken`	(none)

`CallWithBearerToken`é uma permissão da camada de autenticação que autoriza um principal a se autenticar por meio de uma chave de API (token do portador) em vez do AWS SigV4. Ele não mapeia para uma rota. Conceda-o junto com as ações mapeadas de rotas que você deseja que o detentor da chave da API execute.

Acesso ao console

Ação	Rotas autorizadas
`AssumeConsole`	(none)

`AssumeConsole`autoriza um diretor a abrir o console Claude para uma plataforma Claude no espaço de trabalho da AWS por meio do fluxo de federação do console da AWS. Ele não mapeia para uma rota. Conceda-o aos diretores que devem poder clicar em Abrir o console Claude na página de serviço da plataforma Claude na AWS no console da AWS. A chave de `aws-external-anthropic:Capability` condição na `AssumeConsole` ação controla qual recurso do console

(desenvolvedor ou administrador) um diretor pode solicitar. Consulte as políticas do IAM para obter detalhes e como usar o Claude Console para o fluxo de login.

> **Warning**
>
> `aws-external-anthropic:AssumeConsole`emite uma sessão do Claude Console que dura até 12 horas, independente da duração da sessão do próprio chamador. Um chamador cuja sessão do IAM dura menos de 12 horas ainda pode obter uma sessão de console que exceda suas credenciais de origem. Restrinja essa permissão aos diretores que precisam de acesso ao Claude Console e revogue-a imediatamente quando não for mais necessária.

> **Note**
>
> As ações realizadas dentro do Claude Console após a federação não são registradas na AWS CloudTrail nem atribuíveis por meio do IAM. Isso inclui atividades globais no escopo do espaço de trabalho, como a visualização de relatórios de uso. Se você precisar de uma trilha de auditoria para a atividade do console, use os registros de auditoria disponíveis no Claude Console em vez CloudTrail de.

Route-to-action mapeamento

A tabela a seguir lista todas as rotas na plataforma Claude na AWS e a ação do IAM necessária para chamá-la. A ação do IAM da rota estável também autoriza solicitações que usam o `anthropic-beta` cabeçalho. CloudTrail classifica cada ação como um evento de dados (operações de alto volume no plano de dados) ou um evento de gerenciamento (operações no plano de controle).

Método	Rota	Ação do IAM	CloudTrail tipo de evento
`POST`	`/v1/messages`	`CreateInference`	Dados
`POST`	`/v1/messages/count_tokens`	`CountTokens`	Dados
`POST`	`/v1/messages/batches`	`CreateBatchInference`	Dados

Método	Rota	Ação do IAM	CloudTrail tipo de evento
GET	/v1/messages/batches	ListBatchInferences	Dados
GET	/v1/messages/batches/{id}	GetBatchInference	Dados
GET	/v1/messages/batches/{id}/results	GetBatchInference	Dados
POST	/v1/messages/batches/{id}/cancel	CancelBatchInference	Dados
DELETE	/v1/messages/batches/{id}	DeleteBatchInference	Dados
GET	/v1/models	ListModels	Dados
GET	/v1/models/{id}	GetModel	Dados
POST	/v1/files	CreateFile	Dados
GET	/v1/files	ListFiles	Dados
GET	/v1/files/{id}	GetFile	Dados
GET	/v1/files/{id}/content	GetFile	Dados
DELETE	/v1/files/{id}	DeleteFile	Dados
POST	/v1/skills	CreateSkill	Dados
GET	/v1/skills	ListSkills	Dados
GET	/v1/skills/{id}	GetSkill	Dados
DELETE	/v1/skills/{id}	DeleteSkill	Dados
POST	/v1/skills/{id}/versions	UpdateSkill	Dados

Método	Rota	Ação do IAM	CloudTrail tipo de evento
GET	/v1/skills/{id}/versions	GetSkill	Dados
GET	/v1/skills/{id}/versions/{version}	GetSkill	Dados
GET	/v1/skills/{id}/versions/{version}/content	GetSkill	Dados
DELETE	/v1/skills/{id}/versions/{version}	UpdateSkill	Dados
POST	/v1/user_profiles	CreateUserProfile	Dados
GET	/v1/user_profiles	ListUserProfiles	Dados
GET	/v1/user_profiles/{id}	GetUserProfile	Dados
POST	/v1/user_profiles/{id}	UpdateUserProfile	Dados
POST	/v1/organizations/workspaces	CreateWorkspace	Gerenciamento
GET	/v1/organizations/workspaces	ListWorkspaces	Gerenciamento
GET	/v1/organizations/workspaces/{id}	GetWorkspace	Gerenciamento
POST	/v1/organizations/workspaces/{id}	UpdateWorkspace	Gerenciamento
POST	/v1/organizations/workspaces/{id}/archive	ArchiveWorkspace	Gerenciamento

As rotas dos Agentes Gerenciados Claude (agentes, sessões, ambientes, cofres, armazenamentos de memória) seguem o mesmo padrão de rota para ação; para o mapeamento completo por rota, consulte a referência de ações do Anthropic IAM. As rotas do Vault são eventos de gerenciamento; as rotas do agente, da sessão, do ambiente e do armazenamento de memória são eventos de dados.

As rotas que não estão na plataforma Claude na AWS são negadas no gateway por padrão.

Consulte também

- Políticas do IAM, por exemplo, políticas e políticas gerenciadas
- Guia do usuário do AWS IAM para sintaxe de políticas e lógica de avaliação do IAM
- Guia CloudTrail do usuário da AWS para configuração de registros de auditoria

Histórico do documento

A tabela a seguir descreve mudanças importantes na plataforma Claude no guia do usuário da AWS.

Alteração	Descrição	Data
Publicação inicial	Publicou a plataforma Claude no Guia do usuário da AWS.	7 de maio de 2026

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.