aws

Guía del usuario de

Plataforma Claude en AWS

Plataforma Claude en AWS: Guía del usuario de

Table of Contents

¿Qué es la plataforma Claude en AWS?

Acceda a todas las capacidades de la plataforma de Claude a través de AWS con Anthropic-managed infraestructura.

La plataforma Claude en AWS le ofrece la experiencia completa de la plataforma Anthropic, que incluye la API de mensajes, las habilidades de los agentes, la ejecución de código y las funciones beta, a las que puede acceder a través de su cuenta de AWS. A diferencia de Amazon Bedrock, donde AWS opera la pila de inferencias, Anthropic utiliza la plataforma Claude en AWS. AWS proporciona la capa de autenticación (SigV4 o clave de API), el control de IAM-based acceso y la integración de la facturación a través de AWS Marketplace.

> **Note**
>
> Los SDK de Anthropic son compatibles con la plataforma Claude en AWS. Para conocer la disponibilidad de los clientes por idioma, consulte la documentación de los SDK para clientes de Anthropic.

Cómo funciona la integración de la plataforma

Los modelos Claude se ejecutan en la Anthropic-managed infraestructura. Se trata de una integración comercial para la facturación y el acceso a través de AWS. Tanto Anthropic como AWS actúan como procesadores de datos independientes. Los términos de servicio de AWS rigen la plataforma Claude en AWS. También se aplican las condiciones comerciales del servicio, el apéndice de procesamiento de datos y la política de uso de Anthropic.

Tenga en cuenta las siguientes características operativas. Es posible que los datos no residan en AWS. La inferencia puede dirigirse a la nube principal de Anthropic. Los subservicios pueden cambiar sin previo aviso. Defina el `inference_geo` parámetro por solicitud para fijar la inferencia en una zona geográfica específica. Consulte Residencia de datos para obtener más información.

La plataforma Claude en AWS sigue la misma política de retención de datos que la API Claude de origen. La retención cero de datos (ZDR) está disponible bajo petición. Póngase en contacto con su representante de cuentas de Anthropic para habilitarla en su cuenta.

Características de la plataforma Claude en AWS

La plataforma Claude en AWS ofrece las siguientes capacidades:

- Same-day acceso a modelos y funciones: los nuevos modelos y funciones de la API de Claude están disponibles el mismo día en que se lanzan en la API Claude propia.
- Habilidades de agente: utilice las habilidades predefinidas para la generación de documentos (ExcelPowerPoint, Word, PDF) y cree habilidades personalizadas.
- Ejecución de código: ejecute código en el entorno de pruebas gestionado de Anthropic.
- Pensamiento extendido: habilite el pensamiento extendido para tareas de razonamiento complejas.
- Transmisión y procesamiento por lotes: utilice la transmisión SSE en tiempo real o envíe solicitudes por lotes para cargas de trabajo de alto rendimiento.
- Almacenamiento rápido en caché: almacene en caché las herramientas, las indicaciones del sistema y el historial de mensajes para reducir la latencia y los costes.
- API de archivos: carga archivos y haz referencia a ellos en todas las solicitudes.
- Integración de AWS IAM: controle el acceso con políticas de IAM estándar y autenticación SigV4.
- Facturación unificada de AWS: pague a través de su cuenta de AWS existente con una medición basada en el consumo (Marketplace como backend de facturación).

Para ver la lista completa de funciones compatibles, consulte Soporte de funciones.

Cómo está organizada esta guía

En esta guía se tratan los siguientes temas:

- Introducción: configuración de la cuenta, requisitos previos, autenticación e instalación del SDK.
- Uso de la API: modelos disponibles, realización de solicitudes y funciones de soporte.
- Administración de su entorno: residencia de datos, espacios de trabajo, la consola Claude, límites de tarifas y facturación.
- Supervisión y operaciones: CloudTrail registro y seguimiento de las identificaciones de las solicitudes.
- Migración: migración de Amazon Bedrock a la plataforma Claude en AWS.
- Seguridad y control de acceso: referencia a las políticas y acciones de IAM.

La plataforma Claude en AWS frente a Amazon Bedrock

Ambas ofertas le permiten utilizar Claude a través de AWS, pero difieren considerablemente en cuanto a la arquitectura, la superficie de la API y la disponibilidad de las funciones.

	Plataforma Claude en AWS	Amazon Bedrock
¿Quién opera la pila	Anthropic	AWS
Superficie API	API de mensajes antrópicos () `/v1/messages`	API Bedrock (Converse/) InvokeModel
Disponibilidad de funciones	Same-day como Claude API	Depende del cronograma de integración de AWS
Habilidades de los agentes	Disponible	No disponible (requiere la ejecución de código)
Características beta	Compártelo con `anthropic-beta` encabezados (consulta la sección Soporte de funciones)	Requiere el soporte de AWS
Autenticación	AWS IAM/SigV4 o clave de API	AWS IAM/SiGv4 o token portador (solo los SDK de C#, Go y Java)
URL base	`aws-external-anthropic.{region}.api.aws`	`bedrock-runtime.{region}.amazonaws.com`
Cliente SDK	Platform-specific clase de cliente (por ejemplo, `AnthropicAWS` en Python), en versión beta	`AnthropicBedrock` /SDK de Bedrock
Consola	Consola Claude (`platform.claude.com` acceso a través de la consola AWS)	Consola Bedrock

	Plataforma Claude en AWS	Amazon Bedrock
Límites de tarifas y cuotas	Gestionado por Anthropic	Administrado por AWS
Procesadores de datos	Anthropic y AWS	AWS

Si necesita a AWS-operated Claude con la API de Bedrock, consulte Amazon Bedrock.

> **Important**
>
> Anthropic ofrece la plataforma Claude en AWS como oferta de terceros. No está cubierto por los programas de conformidad, las certificaciones o los informes de auditoría estándar de AWS (como la aptitud para las normas SOC, ISO o HIPAA). Los clientes son los únicos responsables de llevar a cabo sus propias diligencias para garantizar que esta oferta de terceros cumpla con sus requisitos normativos, legales y de conformidad.

Cuándo elegir Bedrock: elija Amazon Bedrock si su organización requiere AWS-operated inferencia, AWS como único procesador de datos o cobertura en virtud de los programas de conformidad de AWS (incluidos los requisitos de FedRAMP High, IL4, IL5, SOC, ISO e HIPAA). Bedrock se ejecuta completamente en una AWS-controlled infraestructura con AWS como parte operativa.

Configure su cuenta

La configuración de la plataforma Claude en AWS consta de cuatro fases: registrarse en la consola de AWS, configurar su organización Anthropic, anotar el ID de su espacio de trabajo e iniciar sesión en la consola Claude.

> **Note**
>
> Al registrarse a través de la consola de AWS, se crea una nueva organización de Anthropic vinculada a su cuenta de AWS. Esta organización es independiente de cualquier organización existente que su empresa tenga en Anthropic, incluidas las organizaciones Claude Enterprise adquiridas a través de AWS Marketplace. Las claves de API, los espacios de trabajo y la configuración de Claude Console de una organización propia de Anthropic no se transfieren.
>
> Si ya tienes una oferta privada de Amazon Bedrock, ponte en contacto con tu ejecutivo de cuentas de Anthropic o AWS antes de registrarte para que el descuento se aplique desde tu primera solicitud. Los descuentos no se pueden aplicar retroactivamente al uso realizado antes de que se acepte su oferta privada. Consulta las ofertas privadas.

Paso 1: Inscríbase en la consola de AWS

1. Abra la consola de AWS y vaya a la página de servicios de la plataforma Claude en AWS.
2. Elija Registrarse.
3. Elija Continuar.

La página muestra un banner Sign-up en progreso. Permanezca en la página. Sign-up tarda unos minutos mientras AWS gestiona la suscripción a AWS Marketplace por usted y, a continuación, lo redirecciona automáticamente.

Si su organización tiene una oferta privada de Anthropic, la consola la busca y le pide que la acepte en AWS Marketplace. Consulte Ofertas privadas para obtener más información.

> **Note**
>
> Si utiliza la plataforma Claude en AWS, Anthropic procesa su contenido (como las instrucciones y las respuestas) fuera de AWS. Consulte las políticas de uso de datos de

Anthropic para obtener más información sobre cómo se procesan y almacenan el contenido y los metadatos.

Paso 2: Configura tu organización Anthropic

Una vez que se complete el registro, se te redirigirá a. `platform.claude.com/partner-signup`

1. Introduce la dirección de correo electrónico del propietario de tu organización y selecciona Comenzar.
2. Busca un enlace de configuración en la bandeja de entrada del correo electrónico y síguelo. Si tu navegador muestra la página de inicio de sesión como una cuenta diferente, selecciona Cerrar sesión y continuar.
3. Complete el formulario de detalles de la organización (nombre de la organización, tipo de entidad, país, uso previsto) y elija Completar la configuración.

Al completar la configuración, se crea su organización antrópica. Los términos de servicio de AWS rigen la plataforma Claude en AWS. También se aplican las condiciones comerciales del servicio, el apéndice de procesamiento de datos y la política de uso de Anthropic. La página del servicio de la consola de AWS ahora muestra un panel de navegación a la izquierda con Inicio, claves de API, inicio rápido y espacios de trabajo.

Paso 3: anote el ID de su espacio de trabajo

Un espacio de trabajo predeterminado se aprovisiona automáticamente cuando te registras. Se pueden crear espacios de trabajo adicionales por región; consulte Espacios de trabajo para obtener más información sobre la vinculación regional, la administración de los espacios de trabajo y el alcance de los recursos de IAM.

Busque el ID del espacio de trabajo en Espacios de trabajo en la página de servicios de AWS o en la consola Claude (consulte Uso de la consola Claude). Los ID de los espacios de trabajo utilizan el formato `wrkspc_` seguido de un identificador alfanumérico.

Paso 4: inicie sesión en la consola Claude

El acceso a la consola Claude está federado a través de AWS IAM:

1. Asuma una función de IAM con el permiso. `aws-external-anthropic:AssumeConsole` Consulte las acciones de IAM.
2. En la página de servicios de la plataforma Claude en AWS, seleccione Iniciar sesión. La consola de AWS emite un JWT y lo redirige a. `platform.claude.com`
3. La primera vez que inicie sesión, se le solicitará una dirección de correo electrónico. Introduce tu correo electrónico de trabajo. La plataforma proporciona a su usuario de Claude Console justo a tiempo.

Al iniciar sesión a través de la consola de AWS, la consola Claude se conecta a su organización de la plataforma Claude en AWS. En la barra lateral de Claude Console aparece un indicador de cuenta gestionada por AWS.

Solución de problemas de configuración de cuentas

- «Sign-up fallido: no se pudo activar OutboundWebIdentityFederation»: si ves este cartel rojo la primera vez que lo enviaste, vuelve a seleccionar Continuar. La activación de IAM puede tardar un momento en propagarse.
- No hay indicador de progreso durante el registro: Sign-up tarda unos minutos. La página muestra un banner estático Sign-up en progreso sin barra de progreso mientras AWS aprovisiona su cuenta.
- «Ha iniciado sesión con otra cuenta» después de seguir el enlace de configuración: seleccione Cerrar sesión y continuar. La página te vuelve a autenticar con la dirección de correo electrónico que has introducido.
- Mensaje «No encontrado» durante el inicio de sesión: este mensaje puede aparecer brevemente durante la redirección. Puedes descartarlo.
- La página de uso no muestra ningún dato después de la primera llamada a la API: los datos de uso pueden tardar unos minutos en aparecer en la consola de Claude.

Requisitos previos

Antes de realizar llamadas a la API, asegúrate de tener:

1. Una cuenta de AWS activa con una suscripción a la plataforma Claude en AWS (consulte Configurar su cuenta).
2. La AWS CLI instalada y configurada.
3. La federación de identidades web salientes está habilitada en su cuenta de AWS (un paso de configuración único; consulte Habilitar la federación de identidades web salientes).
4. El ID de su espacio de trabajo (consulte Obtener el ID de su espacio de trabajo).

Habilitar la federación de identidades web salientes

La plataforma Claude en AWS Gateway llama al `sts:GetWebIdentityToken` lado del servidor para acuñar un JWT y lo reenvía a Anthropic. Esta capacidad de STS está inhabilitada de forma predeterminada en todas las cuentas de AWS. Actívala una vez por cuenta:

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

Si la respuesta es`[ERROR] (FeatureEnabled) … already enabled` , la configuración de tu cuenta ya está activada y puedes continuar. Verifica y recupera la URL del emisor de tu cuenta:

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

> **Warning**
>
> Sin este paso, todas las solicitudes se `"Outbound web identity federation is disabled for your account"` devolverán. Este es el error de configuración más común.

Obtén tu ID de espacio de trabajo

Al registrarte, se aprovisiona automáticamente un espacio de trabajo predeterminado en cada región (consulta Configurar tu cuenta). Los espacios de trabajo están enlazados a una sola región de AWS. Puede encontrar el ID del espacio de trabajo en la consola Claude, en Espacios de trabajo, o en la

sección Espacios de trabajo de la página de servicio de la consola de AWS. Consulte los espacios de trabajo para conocer la vinculación de regiones y el alcance de los recursos de IAM.

Configure las variables `ANTHROPIC_AWS_WORKSPACE_ID` y de `AWS_REGION` entorno para que los clientes del SDK las lean automáticamente:

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

La región es obligatoria. El cliente del SDK se activa si no se ha establecido ninguna región. Pase`aws_region`/`awsRegion`al constructor o set `AWS_REGION` (o`AWS_DEFAULT_REGION`). Se admiten todas las regiones comerciales de AWS; las regiones de suscripción requieren que primero active la cuenta en la región.

Autenticación

La plataforma Claude en AWS admite dos métodos de autenticación: AWS IAM con firma de solicitudes SiGv4 (principal) y autenticación con clave de API. Ambos utilizan la misma URL base y el mismo formato de solicitud.

Los SDK de Anthropic (consulte Instalar un SDK) implementan el flujo de autenticación y la resolución de credenciales que se describen a continuación. Si no utilizas un SDK de Anthropic, debes crear SigV4-signed las solicitudes (o presentar tu clave de API como un token portador) en función del punto final regional. `https://aws-external-anthropic.<region>.api.aws` Para obtener SDK-specific información sobre la configuración del cliente y el orden autorizado de resolución de credenciales, consulte la guía de configuración de Claude on AWS en la documentación de Anthropic.

Autenticación de SigV4

SigV4 es la ruta nativa de la empresa y se integra con las políticas, funciones y auditorías de AWS IAM existentes. Configure las credenciales de AWS mediante cualquier método compatible con la cadena de proveedores de credenciales predeterminada de AWS:

- Variables de entorno (`AWS_ACCESS_KEY_ID`,`AWS_SECRET_ACCESS_KEY`,`AWS_SESSION_TOKEN`)
- Archivo de credenciales compartido (`~/.aws/credentials`)
- Archivo de configuración compartido (`~/.aws/config`) que incluye el inicio de sesión único y `credential_process`
- Identidad web (`AWS_WEB_IDENTITY_TOKEN_FILEyAWS_ROLE_ARN`) para IRSA y Actions GitHub
- Credenciales del contenedor de ECS
- Servicio de metadatos de instancias EC2 (IMDS)

Para comprobar que el SDK de Anthropic recoge tus credenciales y las resuelve en el punto de enlace regional correcto, realiza una solicitud de prueba siguiendo los ejemplos de Cómo realizar solicitudes. Una respuesta correcta confirma que las credenciales, la región, la URL base y el ID del espacio de trabajo están configurados correctamente.

Autenticación de la clave de API

Para rutas de integración más sencillas (desarrollo local y scripts), puedes autenticarte con una clave de API en lugar de con SigV4. Configura la variable de `ANTHROPIC_AWS_API_KEY` entorno o pásala `apiKey` al constructor del SDK. El SDK de Anthropic envía la clave como un token portador a la plataforma Claude en el punto final de AWS; IAM autoriza la solicitud mediante la `aws-external-anthropic:CallWithBearerToken` acción (consulte las políticas de IAM).

Genere claves de API en la consola de AWS en la plataforma Claude en AWS → Claves de API. Elija Generar una clave y, a continuación, copie el valor de la clave. Otorgue la acción de `aws-external-anthropic:CallWithBearerToken` IAM a los directores a los que se les debe permitir usar la autenticación con clave de API.

> **Note**
>
> Solo las claves de API creadas en la consola de AWS en la plataforma Claude de AWS funcionan con este servicio.
>
> - Las claves creadas en la consola Claude estándar para el acceso a la API propia no funcionan en la plataforma Claude en el punto final de AWS.
> - Las claves de API de Amazon Bedrock tampoco funcionan: Bedrock utiliza un punto de enlace, un flujo de autenticación y un espacio de nombres de IAM independientes.

Short-term Claves de API

Para los casos en los que desee autenticar la clave de API pero sin la vida útil de una clave de larga duración, Anthropic publica bibliotecas generadoras de tokens que crean claves de API a corto plazo a partir de sus credenciales de AWS IAM. Por defecto, los tokens tienen una vida útil de 12 horas y se pueden asignar a un espacio de trabajo específico y a la acción. `aws-external-anthropic:CallWithBearerToken` Instala el generador para tu idioma, intercambia las credenciales de IAM por una clave de API y pasa la clave al SDK de Anthropic.

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

Short-term Las claves de API aún requieren que el director de IAM de la persona que llama se mantenga en el espacio de trabajo de destino. `aws-external-anthropic:CallWithBearerToken` Caducan por sí solas y no es necesario rotarlas ni revocarlas de forma explícita.

Prioridad de credenciales y resolución regional

La plataforma Claude del SDK de Anthropic en el cliente AWS resuelve las credenciales y la región mediante un orden de prioridad definido. Los nombres de los argumentos siguen las convenciones de cada lenguaje: CamelCase TypeScript para PHP, snake_case para Python y Ruby PascalCase , para Go y patrones de propiedades o constructores para C# y Java.

Para conocer el orden de prioridad autorizado, las variables de entorno compatibles y los argumentos del constructor para cada SDK, consulte Claude sobre la configuración de AWS en la documentación de Anthropic. En general, los argumentos explícitos del constructor tienen prioridad sobre las variables de entorno y tienen `ANTHROPIC_AWS_API_KEY` prioridad sobre la cadena de proveedores de credenciales de AWS predeterminada. La región es obligatoria; a diferencia de `AnthropicBedrock` (que recurre `aus-east-1`), el cliente Claude on AWS aparece si el constructor o`AWS_REGION`/no proporciona ninguna región`AWS_DEFAULT_REGION`.

ID de espacio de trabajo

Cada solicitud de plano de datos debe incluir el ID del espacio de trabajo en el `anthropic-workspace-id` encabezado. Los SDK de Anthropic lo leen de la variable de `ANTHROPIC_AWS_WORKSPACE_ID` entorno de forma predeterminada, o puedes pasar`workspaceId`/`workspace_id`al constructor del cliente. Si utilizas el `Anthropic` cliente base (no el Claude-on-AWS cliente), pasa el encabezado de forma explícita. Consulta Cómo solicitar ejemplos y espacios de trabajo por idioma para saber cómo localizar el ID de tu espacio de trabajo.

Instalar un SDK

Los SDK de cliente de Anthropic son compatibles con la plataforma Claude en AWS. Los siete SDK proporcionan una clase de cliente específica de la plataforma; también puede configurar el cliente base con la URL base de la plataforma Claude en AWS.

Python

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

> **Tip**
>
> En macOS con Homebrew Python u otros entornos Python gestionados externamente, es `pip install` posible que se produzca un error PEP 668`externally-managed-environment`. Primero cree y active un entorno 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**
>
> Los clientes del SDK para la plataforma Claude en AWS están en fase beta.

Modelos disponibles

La plataforma Claude en AWS ofrece el mismo conjunto de modelos Claude que la API Claude original.

Para ver la lista actual de modelos, los identificadores de los modelos, los tamaños de las ventanas de contexto y las capacidades, consulte la descripción general de los modelos en la documentación de Anthropic.

Identificadores de modelo

Los ID de modelo de la plataforma Claude en AWS son idénticos a los que se utilizan en la API Claude de origen (por ejemplo,`claude-opus-4-6`,`claude-sonnet-4-6`,`claude-haiku-4-5`). No hay Bedrock-style ARN ni `anthropic.` prefijos; utilice el ID del modelo exactamente como lo publica Anthropic.

Realizar solicitudes

La plataforma Claude en AWS usa la API de mensajes antrópicos (`/v1/messages`), la misma superficie de API que la plataforma propia de Claude. Las diferencias son la URL base, el método de autenticación y un `anthropic-workspace-id` encabezado obligatorio que identifica el espacio de trabajo al que se dirige la solicitud.

Shell

```
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
```

El cliente lee AWS_REGION (oAWS_DEFAULT_REGION) y ANTHROPIC_AWS_WORKSPACE_ID desde el entorno. Se puede anular pasandoaws_region/awsRegionoworkspace_id/workspaceIdal constructor. Se requieren tanto el identificador de la región como el del espacio de trabajo; el constructor los arroja si alguno de ellos no se puede resolver a partir de estas fuentes.

> **Note**
>
> El `x-amz-security-token` encabezado (curl) solo es necesario para las credenciales temporales, como las funciones de IAM, el SSO o el STS. Omítelo cuando utilice credenciales de usuario de IAM de larga duración. Los clientes del SDK gestionan esto automáticamente en función de la fuente de credenciales.

El `--aws-sigv4` valor sigue el formato. `aws:amz:<region>:<service>` El nombre del servicio SigV4 es `aws-external-anthropic` y la región debe coincidir con la región de la URL de su punto final. Una discrepancia en cualquiera de ellas produce un error genérico de rechazo de firma en lugar de un diagnóstico específico.

Uso del cliente base de Anthropic

Si prefiere no añadir la dependencia del SDK de AWS, puede utilizar el `Anthropic` cliente base. Esta ruta usa los nombres de las variables de entorno estándar del SDK en lugar de los `ANTHROPIC_AWS_*` nombres que lee el cliente dedicado:

```
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'
```

Los SDK de Python y Go se leen `ANTHROPIC_API_KEY` y `ANTHROPIC_BASE_URL` automáticamente; en el caso de los demás lenguajes, se los pasa al constructor. TypeScript En todos los idiomas, pasa el ID del espacio de trabajo al `anthropic-workspace-id` encabezado.

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);
```

Compatibilidad de características

La plataforma Claude en AWS usa directamente la API de Anthropic Messages. Obtendrá una paridad total de funciones con la API Claude propia, excepto cuando se indique en Características que no están disponibles actualmente:

- Funciones beta: pasa el `anthropic-beta` encabezado estándar para acceder a las funciones beta, tal como lo harías con la API Claude.
- Habilidades de agente: usa habilidades de agente prediseñadas y personalizadas con los mismos `container.skills` parámetros y encabezados de la versión beta que la API de Claude. Todas las habilidades predefinidas (ExcelPowerPoint, Word, PDF) funcionan de forma inmediata.
- Ejecución de código: ejecuta código en el entorno de pruebas gestionado por Anthropic utilizando la herramienta de ejecución de código.
- Uso de la herramienta: el uso de la computadora y todas las demás capacidades de uso de herramientas están disponibles.
- Pensamiento extendido: habilite el pensamiento extendido con los mismos parámetros que la API de Claude.
- Transmisión: soporte completo de transmisión SSE para respuestas en tiempo real.
- Procesamiento por lotes: envíe solicitudes por lotes para cargas de trabajo de alto rendimiento.
- Almacenamiento rápido en caché: guarde en caché las herramientas, las indicaciones del sistema y el historial de mensajes para reducir la latencia y los costes. Están disponibles todas las funciones de almacenamiento rápido en caché (TTL de 5 minutos, TTL de 1 hora y almacenamiento en caché automático).
- API de archivos: carga archivos y haz referencia a ellos en todas las solicitudes.
- Etiquetas de espacio de trabajo con integración de IAM: los espacios de trabajo se pueden etiquetar y las etiquetas se transfieren a IAM para el control de acceso basado en atributos y a los informes de costos y uso de AWS para la asignación de costos. El etiquetado de espacios de trabajo es una función de GA de la plataforma Claude en AWS (es una función beta de la API de Claude propia).

Claude Managed Agents

Los agentes gestionados por Claude están disponibles en la plataforma Claude de AWS. Los agentes, las sesiones, los entornos, las bóvedas de credenciales y los almacenes de memoria son

recursos de IAM de primera clase. Consulte Acciones de IAM para ver la referencia de acciones y Uso de la consola Claude para ver las páginas correspondientes. El SDK de Anthropic Agent funciona con la plataforma Claude en AWS sin ninguna integración adicional. Diríjala a la plataforma Claude en la URL base de AWS y autentícate con SigV4 o una clave de API.

Las sesiones autónomas en la plataforma Claude en AWS requieren volver a autenticarse cada 6 horas. Long-running los agentes ejecutados deben actualizar sus credenciales de SigV4 o su clave de API dentro de ese período o la sesión finalizará.

Las siguientes capacidades de Claude Managed Agents no están disponibles actualmente en la plataforma Claude en AWS:

- Resultados: el seguimiento de los resultados de las sesiones de los agentes no está disponible.
- Multi-agent sesiones: las sesiones con varios agentes que interactúan no están disponibles.

Conformidad

> **Important**
>
> Anthropic ofrece este servicio como una oferta de terceros. No está cubierto por los programas de conformidad, las certificaciones o los informes de auditoría estándar de AWS (como la aptitud para las normas SOC, ISO o HIPAA). Los clientes son los únicos responsables de llevar a cabo sus propias diligencias para garantizar que esta oferta de terceros cumpla con sus requisitos normativos, legales y de conformidad. Antes de procesar datos confidenciales o regulados, debe revisar la documentación oficial de cumplimiento de Anthropic a través del Anthropic Trust Center. Consulte los términos de servicio de AWS para obtener más información.

Cómo se modela la membresía a un espacio de trabajo

En la API Claude propia, el acceso se rige por la pertenencia de los usuarios a los espacios de trabajo: un usuario se añade a un espacio de trabajo y hereda los derechos de acceso del espacio de trabajo. La plataforma Claude en AWS reemplaza ese modelo por la autorización de IAM: no hay una lista de usuarios por espacio de trabajo. En cambio, los directores de IAM (usuarios o funciones) tienen políticas que permiten emprender `aws-external-anthropic` acciones contra los ARN

específicos de los espacios de trabajo. La evaluación de la política de IAM determina lo que cada director puede hacer en cada espacio de trabajo.

Conceda y revoque el acceso al espacio de trabajo modificando las políticas de IAM (SCP, límites de permisos, políticas asociadas a la identidad o condiciones a nivel de recursos) en lugar de gestionar la membresía por espacio de trabajo en la consola Claude. Consulte las políticas de IAM para ver ejemplos.

Las funciones no están disponibles actualmente

Las siguientes capacidades no están disponibles actualmente en la plataforma Claude en AWS:

- Preparación para la HIPAA: el HIPAA-ready programa de Anthropic no está disponible en la plataforma Claude de AWS. En su lugar, los clientes que cumplan los requisitos de la HIPAA deberían evaluar a Claude en Amazon Bedrock.
- Niveles de servicio más allá de Standard y Batch: el nivel de prioridad no está disponible.
- API de administración: miembro de la organización, miembro del espacio de trabajo, invitación, clave de API, informe de uso, informe de costos e informe de límites de velocidad: estos puntos de enlace de la API de administración no están disponibles actualmente. El CRUD de Workspace y el cambio de nombre de los puntos de enlace (`/v1/organizations/workspaces`) están disponibles en el espacio de `aws-external-anthropic` nombres; consulte las acciones de IAM. La membresía se modela a través de AWS IAM y no a través de listas de miembros de espacios de trabajo (consulte la sección anterior). El ciclo de vida de las claves de API se administra en la consola de AWS en Claude Platform en AWS → Claves de API. Para obtener datos de uso, costo y límite de velocidad, utilice las vistas de la consola Claude (consulte Monitorización y registro) o la API de uso y costo de Anthropic.
- Límites de gasto: no disponibles. En su lugar, confíe en los controles de facturación de AWS.
- Espacio de trabajo de Claude Code y API de análisis: el espacio de trabajo de Claude Code con límites de velocidad automáticos no está disponible. El uso de Claude Code aparece en la vista de uso general en lugar de en una pantalla dedicada.
- Autenticación OAuth: no se admite. Usa la autenticación mediante clave SigV4 o API.
- OpenAI-compatible Puntos de enlace de la API: no están disponibles en la plataforma Claude de AWS.
- Workspace-level controles geográficos de inferencia: `allowed_inference_geos` y no `default_inference_geo` están disponibles. `inference_geo`En su lugar, establézcalos en cada solicitud.

Residencia de datos

La plataforma Claude en AWS admite las siguientes geografías de inferencia:

- EE. UU.: La inferencia se realiza en los centros de datos estadounidenses. Se aplica un multiplicador de precios de 1,1 veces.
- Global: la inferencia puede dirigirse a cualquier centro de Anthropic-operated datos del mundo. Se aplica un precio estándar.

Defina la geografía de inferencia por solicitud con el `inference_geo` parámetro:

> **Note**
>
> El `inference_geo` parámetro es compatible con los modelos Claude Opus 4.6, Claude Sonnet 4.6 y posteriores. Las solicitudes con `inference_geo` Claude Opus 4.5, Claude Sonnet 4.5 o Claude Haiku 4.5 devuelven un error 400. Consulte Residencia de datos para obtener detalles sobre la disponibilidad del modelo.

Shell

```
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
```

Si lo omite`inference_geo`, el valor predeterminado de la solicitud es. `global`

Workspace-level los controles geográficos de inferencia (`allowed_inference_geos`y`default_inference_geo`) no están disponibles en la plataforma Claude en AWS. En su lugar`inference_geo`, establézcalo en cada solicitud.

Espacios de trabajo

Las solicitudes de inferencias y recursos en la plataforma Claude en AWS se dirigen a un espacio de trabajo. El ID del espacio de trabajo se introduce en el `anthropic-workspace-id` encabezado de estas llamadas a la API. Los ID de los espacios de trabajo utilizan el formato etiquetado `wrkspc_` seguido de un identificador alfanumérico (por ejemplo,`wrkspc_01AbCdEf23GhIj`). Consulta Cómo obtener el ID de tu espacio de trabajo si aún no lo tienes.

Alcance del espacio de trabajo

Los espacios de trabajo están enlazados a una sola región de AWS. Solo se `us-west-2` puede acceder a un espacio de trabajo creado en el `us-west-2` punto final. El uso, las cuotas, el coste, los archivos, los lotes y las habilidades se acumulan por espacio de trabajo, lo que te proporciona un desglose por región en la consola de Claude.

Los espacios de trabajo también son el principal recurso de IAM para Claude Platform en AWS. Usted concede o deniega el acceso a espacios de trabajo específicos mediante las políticas de AWS IAM mediante el ARN del espacio de trabajo. El segmento de recursos del ARN es el mismo identificador con el `wrkspc_` prefijo que se introduce en el encabezado: `anthropic-workspace-id`

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

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

Consulte las políticas de IAM para ver ejemplos de políticas.

Creación de espacios de trabajo

Un espacio de trabajo predeterminado se aprovisiona automáticamente al registrarse. Se pueden crear, actualizar, cambiar el nombre y archivar espacios de trabajo adicionales a través de los `/v1/organizations/workspaces` puntos finales, con la autorización de `aws-external-anthropic` las acciones correspondientes (, `CreateWorkspaceUpdateWorkspace,ArchiveWorkspace`); consulte Acciones de IAM.

Configurar el ID del espacio de trabajo en su cliente

Cada solicitud de plano de datos debe incluir el ID del espacio de trabajo en el `anthropic-workspace-id` encabezado. Cuando utilizas un Claude-on-AWS cliente del SDK de Anthropic, hay tres formas de suministrarlo:

1. Variable de entorno: `ANTHROPIC_AWS_WORKSPACE_ID` se establece y el cliente la lee automáticamente.

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

2. Argumento constructor: pass`workspaceId`/`workspace_id`(el nombre sigue la convención de lenguaje del SDK) al crear una instancia del cliente.
3. Per-request anular: pasa el encabezado directamente a una solicitud individual si necesitas abordar varios espacios de trabajo del mismo cliente.

Si usa el `Anthropic` cliente base (sin el backend de AWS), configure el `anthropic-workspace-id` encabezado de forma explícita en cada solicitud; consulte Ejemplos de solicitudes por idioma. Para ver los nombres de los SDK-specific argumentos, consulte la documentación de los SDK de Anthropic Client.

Etiquetado de espacios de trabajo

Las etiquetas de espacio de trabajo son pares clave-valor adjuntos a un espacio de trabajo. Se integran con AWS IAM para el control de acceso basado en atributos y con los informes de costo y uso de AWS para la asignación de costos (consulte Supervisión y registro para obtener información sobre CUR). El etiquetado es GA en la plataforma Claude de AWS.

Actualice las etiquetas del espacio de trabajo existente con `TagResource` y `UntagResource` en el espacio de `aws-external-anthropic` nombres. Las mismas claves de etiquetas pueden impulsar las condiciones de IAM y la asignación de costes sin necesidad de configuración adicional.

Tag-based control de acceso

Puede bloquear `aws-external-anthropic` las acciones en los valores de las etiquetas del espacio de trabajo mediante la clave de contexto de `aws:ResourceTag/<key>` condición.
La siguiente política permite la inferencia solo en los espacios de trabajo etiquetados: `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"
        }
      }
    }
  ]
}
```

Usa `aws:RequestTag/<key>` y `aws:TagKeys` condiciona las claves `TagResource` para hacer cumplir las políticas de etiquetado (por ejemplo, exigiendo que el espacio de trabajo lleve `CostCenter` etiquetas). `Owner` Consulte las políticas de IAM para ver más ejemplos.

Uso de la consola Claude

La plataforma Claude en AWS usa la consola Claude estándar en platform.claude.com. Al iniciar sesión desde la consola de AWS, aparece un indicador de cuenta gestionada por AWS en la barra lateral de Claude Console. La consola abarca su organización de la plataforma Claude en AWS. Proporciona análisis de uso, desgloses de costos, visibilidad con límites de velocidad, visibilidad del espacio de trabajo y páginas para administrar archivos, habilidades de los agentes, trabajos por lotes, agentes, sesiones, entornos, almacenes de credenciales y almacenes de memoria.

Inicio de sesión

El acceso a la consola Claude está federado a través de AWS IAM. Consulte Configurar su cuenta para ver el flujo completo de inicio de sesión por primera vez. En resumen:

1. Asuma una función de IAM con el `aws-external-anthropic:AssumeConsole` permiso. Consulte las acciones de IAM.
2. Diríjase a la página de la plataforma Claude en AWS en la consola de AWS.
3. Elija Open Claude Console. La consola de AWS emite un JWT y lo redirige a. `platform.claude.com`
4. La primera vez que inicie sesión, se le solicitará una dirección de correo electrónico; introduzca su correo electrónico de trabajo. La plataforma proporciona a su usuario de Claude Console justo a tiempo.

Hay dos funciones disponibles en Claude Console: administrador y desarrollador. El rol de administrador otorga acceso a todas las páginas y configuraciones de la consola Claude disponibles para la plataforma Claude en AWS. El rol de desarrollador otorga acceso de lectura a la información de uso, costo, límite de velocidad y espacio de trabajo. Ponte en contacto con tu representante de cuentas de Anthropic para asignar el rol de administrador o desarrollador a un director.

Páginas disponibles

La columna Via AWS Gateway indica si la página lee y escribe datos a través de la puerta de enlace de AWS (y, por lo tanto, se rige por las acciones de IAM). Las páginas marcadas como No leen los metadatos a nivel de la organización directamente a través de las API de Anthropic y omiten las comprobaciones de acción de IAM.

Page	Disponible	A través de AWS Gateway	Notas
Uso	Sí	No	Vea el uso de los tokens por modelo, espacio de trabajo y dimensión. Los datos pueden tardar unos minutos en aparecer después de una solicitud.
Costo	Sí	No	Vea los desgloses de costos por modelo y espacio de trabajo. AWS Cost Explorer muestra la partida de CCU agregada.
Límites	Sí	No	Vea los límites de velocidad (solo lectura).
Espacios de trabajo	Sí	No	Vea los espacios de trabajo por región (solo lectura).
Archivos	Sí	Sí	Vea y administre los archivos cargados.
Habilidades	Sí	Sí	Vea y gestione las habilidades de los agentes.
Lotes	Sí	Sí	Vea y gestione los trabajos de procesamiento por lotes.
Agentes	Sí	Sí	Vea y gestione las definiciones de los agentes.
Sessions (Sesiones)	Sí	Sí	Vea las sesiones de los agentes y el historial de eventos.
Entornos	Sí	Sí	Vea y gestione las configuraciones de los contenedores en la nube para las sesiones.
Bóvedas de credenciales	Sí	Sí	Vea y administre los almacenes de credencia les para la autenticación de sesiones.
Almacenes de memoria	Sí	Sí	Vea y gestione la memoria de los agentes persistentes.

Page	Disponible	A través de AWS Gateway	Notas
Claves de API	No	N/A	Administre las claves de API en la consola de AWS (Plataforma Claude en AWS → Claves de API). Consulte Autenticación.
Miembros	No	N/A	No se usa. AWS IAM administra el acceso.
Facturación	No	N/A	No se usa. AWS Marketplace gestiona la facturación y la facturación. Consulte el desglose de los costos en la página de costos.
Código Claude	No	N/A	Vea el uso del código Claude en la página de uso.

Cambiando de organización

La consola Claude no admite el cambio de organización a la plataforma Claude en AWS. Para acceder a otra organización, cierre sesión y vuelva a autenticarse a través de la consola de AWS con el rol de IAM de la cuenta de AWS de esa organización.

Límites de tarifas y cuotas

La plataforma Claude en AWS asigna límites de velocidad de nivel 1 al suscribirse. Anthropic gestiona los límites de tarifas directamente, no a través de los sistemas de cuotas de AWS.

Límites predeterminados

La plataforma Claude en AWS utiliza el programa de niveles estándar de Anthropic, idéntico a la API Claude de origen. Se aplican límites de nivel 1 por espacio de trabajo. Los límites se agrupan por familia de modelos. Por ejemplo, todos los modelos Opus comparten un límite combinado, todos los modelos Sonnet comparten otro y todos los modelos Haiku comparten un tercero.

Para conocer los valores de nivel 1 actuales (RPM, ITPM, OTPM) y los umbrales de niveles superiores, consulte los límites de velocidad en el sitio web de documentación de Anthropic. La página de Anthropic es la fuente de información y se actualiza cuando cambian los límites.

Cabeceras con límite de velocidad

Todas las respuestas incluyen encabezados que indican el estado actual de tu límite de velocidad. Encabezados clave:

- `anthropic-ratelimit-requests-limit`— Número máximo de solicitudes por minuto
- `anthropic-ratelimit-requests-remaining`— Solicitudes que permanecen en la ventana actual
- `anthropic-ratelimit-requests-reset`— Hora en la que se restablece el límite de solicitudes (RFC 3339)
- `anthropic-ratelimit-tokens-limit`— Número máximo de fichas combinadas (entrada +salida) por minuto
- `anthropic-ratelimit-tokens-remaining`— Las fichas combinadas que quedan en la ventana actual
- `anthropic-ratelimit-tokens-reset`— Hora en la que se restablece el límite de fichas combinadas (RFC 3339)
- `anthropic-ratelimit-input-tokens-limit`//— encabezados `-remaining` `-reset` Input-token-specific

- `anthropic-ratelimit-output-tokens-limit/-remaining/-reset`— encabezados Output-token-specific
- `retry-after`— En una respuesta de 429, el número de segundos que hay que esperar antes de volver a intentarlo

Consulte los encabezados de respuesta en el sitio web de documentación de Anthropic para ver el conjunto completo.

¿Solicita límites más altos

A diferencia de la API Claude propia, el avance automático de nivel no se aplica a la plataforma Claude en AWS. Para solicitar límites más altos, póngase en contacto con el representante de su cuenta de Anthropic e indique su ID de espacio de trabajo y el rendimiento deseado. Para ver los umbrales de nivel y otros detalles, consulta los límites de velocidad en el sitio web de documentación de Anthropic.

Errores en el límite de velocidad

Cuando superas un límite de velocidad, la API devuelve HTTP 429 con un `rate_limit_error` tipo. Implementa un retroceso exponencial con fluctuación en tu lógica de reintento. El `retry-after` encabezado indica cuántos segundos hay que esperar antes de volver a intentarlo.

Facturación

Claude Platform en AWS utiliza AWS Marketplace como backend de facturación para la medición basada en el consumo. Anthropic mide el uso cada hora a través de AWS Marketplace y AWS emite facturas mensuales a partir de su factura de AWS actual.

La facturación se realiza únicamente con atrasos (usted paga por lo que ha utilizado) y antes de impuestos (AWS aplica la configuración fiscal de su cuenta).

El uso se expresa en unidades de consumo (CCU) de Claude y cuesta 0,01 USD por CCU. El precio de la CCU es fijo y no incluye descuentos. Anthropic valora el uso de tus fichas en USD según las tarifas estándar por modelo y por función, aplica cualquier descuento negociado y, a continuación, convierte el resultado en CCU a 0,01 USD por CCU. Los descuentos se traducen en un menor número de CCU medidas, no en un precio de CCU más bajo. Las CCU no son créditos prepagos; no hay saldo ni compromiso de CCU. Consulte los precios para conocer la definición de CCU y las tarifas de los tokens por modelo.

Supervisión y registro

AWS CloudTrail puede capturar todas las solicitudes a la plataforma Claude en AWS. De forma predeterminada, las operaciones del espacio de trabajo y el almacén se registran como eventos de administración. Todas las demás operaciones (inferencia, lote, archivo, habilidad, modelo, perfil de usuario y Claude Managed Agents, excepto los almacenes) son eventos de datos. Registrarlos requiere una configuración explícita y conlleva cargos adicionales. CloudTrail Consulte las acciones de IAM para ver la clasificación completa de los tipos de eventos y la CloudTrail documentación de AWS para obtener detalles de configuración.

Al configurar el registro CloudTrail de eventos de datos, seleccione el tipo `AWS::AWSExternalAnthropic::Workspace` de recurso. Este es el tipo de CloudTrail recurso para la plataforma Claude en los espacios de trabajo de AWS y es necesario para capturar eventos del plano de datos (operaciones de inferencia, por lotes, de archivos y otras operaciones por espacio de trabajo). Limite los selectores de eventos de datos a los ARN específicos del espacio de trabajo si desea registrar la actividad de un subconjunto de espacios de trabajo y no de toda la cuenta.

ID de solicitud

Cada respuesta incluye dos ID de solicitud en los encabezados de respuesta:

- ID de solicitud de AWS (**`x-amzn-requestid`**): el ID principal, indexado en CloudTrail. Úselo cuando investigue las solicitudes a través de las herramientas de AWS o cuando se ponga en contacto con el soporte de AWS.
- ID de solicitud antrópica (**`request-id`**): el ID secundario. Úselo cuando se comunique con el soporte de 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

Actualmente, el SDK de Ruby no muestra un descriptor de acceso de respuesta sin procesar, por lo que los ID de las solicitudes no se pueden leer en los encabezados de respuesta de Ruby. Si necesitas registrar los ID de solicitud, usa cualquiera de los lenguajes anteriores o captura los ID en la capa de proxy HTTP.

Anthropic recomienda registrar tu actividad al menos 30 días seguidos para comprender los patrones de uso e investigar cualquier posible problema.

> **Note**
>
> AWS CloudTrail se configura en su cuenta de AWS. Habilitar el registro no proporciona a AWS o Anthropic acceso a su contenido más allá de lo necesario para la facturación y el funcionamiento del servicio.

Métricas de uso y paneles de control

La plataforma Claude en AWS no publica métricas de uso por espacio de trabajo en Amazon CloudWatch. En cambio, los recuentos de fichas, el volumen de solicitudes y el uso por modelo están disponibles en las secciones de uso de Anthropic:

- Vistas de uso de la consola Claude: cuando un director se federa en la consola Claude con `aws-external-anthropic:AssumeConsole` (consulte las políticas de IAM), los paneles de uso muestran el volumen de solicitudes, el recuento de tokens y los desgloses por modelo para los espacios de trabajo accesibles. Account-wide las vistas de uso requieren la capacidad de la consola de administración.
- API de uso y costo de Anthropic: para obtener acceso programático a los datos de uso y costo, incluida la agregación por espacio de trabajo y por modelo, consulte la API de uso y costo en la documentación de Anthropic.

Para la supervisión operativa (tasas de error, latencia, encabezados de límite de velocidad), utilice los encabezados de respuesta que se muestran en cada solicitud (consulte Límites y cuotas de velocidad) junto con los ID de solicitud de AWS capturados en ellos. CloudTrail

Asignación y monitoreo de costos

Los cargos de Claude Platform on AWS aparecen en su factura de AWS en el servicio Claude Platform on AWS, con las dimensiones de uso por modelo. Use el informe de costo y uso (CUR) de AWS combinado con etiquetas de espacio de trabajo para una asignación de costos detallada:

1. Etiquete sus espacios de trabajo con las dimensiones de asignación que más le interesen (equipo, aplicación, entorno, centro de costes). Consulte Espacios de trabajo para obtener detalles sobre el etiquetado.

2. En la consola de facturación de AWS, active cada clave de etiqueta como etiqueta de asignación de costes. Tras la activación, el uso realizado en la fecha de activación o después se divide por etiqueta en el CUR.
3. En CUR o AWS Cost Explorer, agrupe por `resourceTags/user:<tag-key>` columnas para desglosar el gasto por etiqueta de espacio de trabajo.

Las etiquetas del espacio de trabajo fluyen hasta las líneas de CUR para todo el uso de Claude Platform en AWS. Las mismas claves de etiquetas impulsan tanto el control de IAM-based acceso como la asignación de costos. Consulte las etiquetas de asignación de costes en la documentación de facturación de AWS para conocer los pasos de configuración y los retrasos en la activación.

Migración desde Amazon Bedrock

Si actualmente usa Claude en Bedrock, la migración a Claude Platform en AWS requiere cambios en toda la integración. La firma SigV4 sigue siendo compatible, pero el contexto de firma, la URL base, el formato de la API, los ID de modelo, el cliente y el paquete del SDK, el formato de transmisión, los encabezados de las solicitudes y la disponibilidad regional cambian. En la siguiente tabla se resumen las diferencias.

¿Qué cambios

	Amazon Bedrock	Plataforma Claude en AWS
URL base	`bedrock-runtime.{region}.amazonaws.com`	`aws-external-anthropic.{region}.api.aws`
Formato de API	Bedrock Converse/ InvokeModel	API de mensajes antrópicos () `/v1/messages`
ID de modelo	`anthropic.claude-opus-4-7-v1`	`claude-opus-4-7`
Cliente SDK	`AnthropicBedrock` /SDK de Bedrock	Platform-specific cliente (`AnthropicAWS` ,`AnthropicAws` ,`AnthropicAwsClient` , etc.), en versión beta
paquete SDK	`anthropic[bedrock]` `@anthropic-ai/bedrock-sdk`, etc.	`anthropic[aws]` `@anthropic-ai/aws-sdk` , etc. (consulte Instalar un SDK)
Nombre del servicio SiGv4	`bedrock`	`aws-external-anthropic`
Formato de streaming	AWS EventStream	SSE (igual que Claude API)

	Amazon Bedrock	Plataforma Claude en AWS
encabezado del espacio de trabajo	No aplicable	`anthropic-workspace-id` obligatorio
Disponibilidad por región	Varias regiones comerciales	Todas las regiones comerciales de AWS (las regiones de suscripción obligatoria requieren la suscripción a la cuenta)

Lo que gana

- Por lo general, el acceso a los nuevos modelos y funciones se realiza el mismo día, sin necesidad de un paso independiente para la integración de los socios
- Habilidades de agente para la generación de documentos (ExcelPowerPoint, Word, PDF)
- Ejecución de código en el entorno sandbox gestionado por Anthropic
- Funciones beta a través del `anthropic-beta` encabezado (consulte Funciones que no están disponibles actualmente)
- Claude Console para analizar la visibilidad de las cuotas y el uso
- Soporte antrópico directo
- La autenticación mediante clave API como alternativa a SiGv4 (consulte Autenticación)

¿Qué permanece igual

- Autenticación de AWS IAM (SiGv4)
- Facturación a través de su cuenta de AWS
- Retirada del compromiso de AWS

Dificultades relacionadas con la migración

> **Warning**
>
> Active primero la federación de identidades web salientes. Si su cuenta de AWS no ha utilizado anteriormente la plataforma Claude en AWS, debe habilitar la federación de identidades web salientes una vez por cuenta antes de realizar solicitudes. Sin este paso, todas las solicitudes fallan y se produce un error de federación. Este paso no es obligatorio para Bedrock.

> **Warning**
>
> La retención cero de datos (ZDR) es opcional en la plataforma Claude en AWS. En Bedrock, AWS es el procesador de datos y Anthropic no conserva las entradas o salidas de inferencias; el programa ZDR de Anthropic no se aplica allí. En Claude Platform en AWS, Anthropic es el procesador de datos y ZDR sigue el modelo de API Claude propio: está disponible previa solicitud a través del representante de su cuenta de Anthropic. Confirme la inscripción en ZDR antes de migrar las cargas de trabajo de producción que dependen de las garantías de retención de datos.

Consideraciones comerciales

- Condiciones de servicio: las condiciones de servicio de AWS rigen la plataforma Claude en AWS. También se aplican las condiciones comerciales del servicio, el apéndice de procesamiento de datos y la política de uso de Anthropic.
- Descuentos y ofertas privadas: los descuentos negociados y las ofertas privadas de AWS Marketplace no se transfieren automáticamente entre Bedrock y Claude Platform en AWS. Trabaje con su representante de cuentas de Anthropic para configurar las condiciones comerciales de la plataforma Claude en AWS.

Políticas de IAM

La plataforma Claude en AWS se integra con AWS IAM para el control de acceso. Usted concede o deniega el acceso a acciones específicas de la API en espacios de trabajo específicos mediante la sintaxis de políticas de IAM estándar.

El nombre del servicio SigV4 y el espacio de nombres de la acción de IAM son. `aws-external-anthropic` Las acciones siguen el patrón `aws-external-anthropic:<Action>` (por ejemplo,). `aws-external-anthropic:CreateInference`

Ejemplo: denegar la inferencia por lotes

La siguiente política permite la inferencia en tiempo real y, al mismo tiempo, bloquea el procesamiento por lotes, un requisito común para ZDR-sensitive las cargas de trabajo:

```
{
  "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": "*"
    }
  ]
```

```
}
```

La `GetBatchInference` acción autoriza tanto la ruta de los metadatos del lote como la ruta de los resultados del lote. Al negarlo, además`ListBatchInferences`, se bloquean tanto las lecturas como la enumeración por lotes.

La `Allow` declaración enumera `List*` acciones específicas en lugar `Get*` de utilizar caracteres comodín. Los caracteres comodín concederían lecturas `GetFile` (lo que descarga los bytes del archivo) y otras lecturas no deseadas; en cualquier caso, las `Deny` anularía, `Allow` pero la forma explícita es el patrón más seguro de modelar.

Ejemplo: inferencia sincrónica en un único espacio de trabajo

Otorga los permisos mínimos a un director de IAM que realice inferencias en un espacio de trabajo de producción:

```
{
  "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**
>
> El `List*` comodín de esta política también coincide`ListWorkspaces`, ya que depende del ámbito de la cuenta. La restricción ARN del espacio de trabajo lo filtra silenciosamente, por lo que esta política no autoriza la publicación de espacios de trabajo. Si su cuenta de servicio necesita enumerar los espacios de trabajo, añada una declaración separada para with. `Allow` `ListWorkspaces` `Resource: "*"`

Esta política asume la autenticación SigV4 de AWS. Si el principal se autentica con una clave de API, también la concede `aws-external-anthropic:CallWithBearerToken` (consulte Autenticación).

Ejemplo: aislamiento del espacio de trabajo por cliente

Restringe un rol a un solo espacio de trabajo:

```
{
  "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

El `aws-external-anthropic:*` comodín de la primera declaración incluye acciones relacionadas con la cuenta (`CreateWorkspace`,`ListWorkspaces`) que la restricción ARN del espacio de trabajo filtra silenciosamente. Esto es coherente con la intención de «aislar» (el rol no puede crear ni enumerar espacios de trabajo), pero la política contiene permisos que no tienen ningún efecto. Consulte la automatización del aprovisionamiento para conocer el patrón centrado en la cuenta.

La segunda declaración concede `CallWithBearerToken` y `AssumeConsole` en todos los recursos porque ambas son acciones sin ruta que no se vinculan a un ARN del espacio de

trabajo. Omita la segunda afirmación si el rol solo usa SigV4 y nunca se federa en la consola Claude.

Ejemplo: bloqueo de funciones para un espacio de trabajo ZDR-sensitive

Bloquea el procesamiento por lotes y la carga de archivos en un espacio de trabajo específico y deja disponible la inferencia sincrónica. Resulta útil cuando un espacio de trabajo gestiona datos con cero retención de datos (ZDR) que no deben permanecer en el servidor. Adjunta esta política a una política de autorización, como `AnthropicLimitedAccess` la que se muestra arriba en el ejemplo de espacio de trabajo único; por sí sola, una Deny-only política no concede permisos:

```
{
  "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

Esta denegación solo bloquea la creación. Las demás acciones de archivo y lote no se rechazan a menos que también las enumere. Para un bloqueo total en el que el espacio de trabajo nunca deba almacenar archivos o lotes, deniegue también las `aws-external-anthropic:GetFile` opciones`aws-external-anthropic:ListFiles`,`aws-external-anthropic:DeleteFile`,`aws-external-anthropic:GetBatchInference`, `aws-external-anthropic:ListBatchInferencesaws-`

`external-anthropic:CancelBatchInference`, y. `aws-external-anthropic:DeleteBatchInference`

Ejemplo: automatización del aprovisionamiento

Otorga a un CI/CD rol las acciones necesarias para crear y administrar espacios de trabajo, sin ningún permiso de inferencia:

```
{
  "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": "*"
    }
  ]
}
```

`CreateWorkspacey` `ListWorkspaces` son operaciones que se centran en la cuenta. Especificar un ARN de espacio de trabajo en estas acciones no tiene ningún efecto; utilícelo. `Resource: "*"`

Políticas administradas

AWS proporciona políticas administradas para patrones de acceso comunes:

- **`AnthropicFullAccess`**: Subvenciones `aws-external-anthropic:*` para todos los recursos.
- **`AnthropicReadOnlyAccess`**: Subvenciones `Get*` `List*` y `CallWithBearerToken` sobre todos los recursos.
- **`AnthropicInferenceAccess`**: Otorga las ReadOnly acciones más las acciones de inferencia (`CreateInference`,`CreateBatchInference`,`CancelBatchInference`,`DeleteBatchInference`,`C` en todos los recursos.

- **`AnthropicLimitedAccess`**: Otorga las `AnthropicInferenceAccess` acciones y todas las acciones de Claude Managed Agents (agentes, sesiones, entornos, bóvedas, almacenes de memoria) en todos los recursos.
- **`AnthropicSelfHostedEnvironmentAccess`**: Otorga los permisos que un trabajador de entorno aislado autohospedado necesita para sondear y procesar los elementos de trabajo del entorno, leer el entorno, la sesión y los recursos de habilidades asociados y actualizar el estado de la sesión. Adjunte esta política a la función de IAM que asume su trabajador del entorno autohospedado. `AnthropicSelfHostedEnvironmentAccesses` el rol único predeterminado recomendado; si ejecuta el sondeo de trabajo y el entorno limitado por sesión bajo principios de IAM independientes, puede dividir estos permisos en dos políticas personalizadas más específicas para obtener los privilegios mínimos.

`AnthropicInferenceAccesses` la política gestionada más limitada y suficiente para realizar inferencias. Mediante los `List*` caracteres comodín `Get*` y, otorga acceso de lectura a todos los recursos de la API del espacio de nombres, incluidos los contenidos de los archivos descargados y de la memoria. `GetFile` `GetMemoryStore` No permite la creación o eliminación de archivos o habilidades, la administración de perfiles de usuario, la mutación del espacio de trabajo ni la federación de consolas.

> **Note**
>
> `AnthropicReadOnlyAccessAnthropicInferenceAccess`, y no `AnthropicLimitedAccess` lo concedas`AssumeConsole`. Los directores que necesiten unirse a la Claude Console necesitan una subvención independiente, ya sea mediante una política personalizada `AnthropicFullAccess` o a través de una política personalizada. `aws-external-anthropic:AssumeConsole` Consulte Cómo federarse a la consola Claude.

> **Note**
>
> `CreateInferencey` `CreateBatchInference` son acciones independientes. Negar una no bloquea la otra. Si pretende impedir todas las llamadas modelo, rechace ambas.

¿Se está federando a la consola Claude

`aws-external-anthropic:AssumeConsole`permite a un director de IAM federarse en la Anthropic-operated consola Claude. El acceso a la consola sigue rigiéndose por la IAM en la mayoría de las operaciones, pero un subconjunto de las operaciones de administración (principalmente las vistas de uso que no tienen la API de IAM correspondiente) están limitadas a la capacidad con la que se federó el principal.

Existen dos capacidades:

- **`developer`**— permite las operaciones que un desarrollador de la plataforma Claude en AWS necesita para su trabajo diario: ejecutar inferencias desde la consola, leer los datos del espacio de trabajo y ver el uso personal.
- **`admin`**— además, permite las operaciones de la consola solo para administradores, incluidas las vistas de uso de toda la cuenta y los ajustes administrativos que no aparecen en las acciones de IAM.

Controle qué capacidad puede solicitar un director con la clave de condición de la `aws-external-anthropic:Capability` acción: `AssumeConsole`

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

`AnthropicFullAccess`subvenciones `AssumeConsole` sin restricción de capacidad. Para cualquier concesión más limitada (acceso a la consola solo para desarrolladores o acceso solo para

administradores), adjunta una política personalizada que cumpla con la condición, tal y como se muestra arriba. `AssumeConsole` `aws-external-anthropic:Capability`

Llamar con fichas al portador

`aws-external-anthropic:CallWithBearerToken`autoriza la ruta de SigV4-free solicitud utilizada cuando una clave de API se presenta como un token portador. Cualquier director que se autentique con una clave de API necesita esta acción en el espacio de trabajo de destino. Esto se aplica tanto si utilizas el `Authorization: Bearer` encabezado como `ANTHROPIC_AWS_API_KEY` si lo configuras directamente.

Esto es obligatorio para las personas que llaman a la clave de la API, además de las acciones de inferencia (`CreateInference`,`CreateBatchInference`, etc.). Sin ellas`CallWithBearerToken`, las solicitudes de claves de API se rechazan antes de pasar a la comprobación de autorización de la inferencia. Las personas que llaman a SigV4 no necesitan esta acción.

`AnthropicReadOnlyAccess`, `AnthropicInferenceAccessAnthropicLimitedAccess`, y `AnthropicFullAccess` todas incluyen. `CallWithBearerToken` Si escribes una política personalizada para el acceso a la clave de la API, agrégala de forma explícita.

Federación de identidades web salientes (necesaria para el acceso a la consola)

La consola Claude se ejecuta en una infraestructura antrópica, no en AWS. Cuando un director de IAM llama`AssumeConsole`, AWS STS emite un token de identidad web dirigido a la audiencia de Anthropic; a continuación, la consola de Anthropic acepta ese token y establece la sesión federada. Para que esto funcione, la cuenta de AWS debe permitir la federación de identidades web salientes para la `aws-external-anthropic` audiencia.

Los directores que llamen `AssumeConsole` necesitan los siguientes permisos de STS además de la `aws-external-anthropic:AssumeConsole` acción:

- **`sts:GetWebIdentityToken`**— permite emitir el token de identidad web que consume la consola Anthropic.
- **`sts:TagGetWebIdentityToken`**— permite adjuntar etiquetas de sesión al token de identidad web. La plataforma Claude en AWS usa estas etiquetas para transmitir la capacidad del director y el contexto del espacio de trabajo a la consola.

Incluya tanto en la política de confianza de cualquier función que asuman los usuarios de la consola como en la inline/managed política asociada a las identidades de los usuarios a los que llaman `AssumeConsole` directamente. `AnthropicFullAccess`incluye ambas acciones de STS. Los entornos que deniegan `sts:*` a nivel de SCP o límite de permisos deben permitir explícitamente estas dos acciones para que la federación de consolas se lleve a cabo correctamente.

Acciones de IAM para la plataforma Claude en AWS

Referencia de acción de IAM para controlar el acceso a la plataforma Claude en AWS a través de las políticas de AWS.

La plataforma Claude en AWS utiliza AWS IAM para el control de acceso. Cada ruta de la API se asigna a una acción de IAM en el `aws-external-anthropic` espacio de nombres. En esta página se enumeran todas las acciones, las rutas que cada acción autoriza y las políticas gestionadas disponibles para los patrones de acceso comunes. Para obtener información sobre la configuración y la autenticación de la plataforma, consulte ¿Qué es la plataforma Claude en AWS? .

Detalles del servicio

prefijo del servicio IAM	`aws-external-anthropic`
Tipos de recurso	`workspace`
ARN del espacio de trabajo	`arn:aws:aws-external-anthropic:{region}:{account-id}:workspace/{workspace-id}`

La región ARN siempre se rellena y coincide con la región a la que está enlazado el espacio de trabajo. El segmento de recursos es el identificador del espacio de trabajo etiquetado (`wrkspc_…`), el mismo valor que se introduce en el `anthropic-workspace-id` encabezado.

Acciones

El servicio define 65 acciones en 15 grupos. Las acciones siguen la `VerbNoun` convención de AWS y utilizan la disciplina verbal para que`Get*`, junto con los `List*` comodines, se establezca un límite limpio de solo lectura.

Inferencia

Action	Rutas autorizadas
`CreateInference`	`POST /v1/messages`

Action	Rutas autorizadas
CountTokens	POST /v1/messages/count_tokens

Procesamiento por lotes

Action	Rutas 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

Action	Rutas autorizadas
GetModel	GET /v1/models/{id}
ListModels	GET /v1/models

Archivos

Action	Rutas autorizadas
CreateFile	POST /v1/files
GetFile	GET /v1/files/{id} GET /v1/files/{id}/content
ListFiles	GET /v1/files

Action	Rutas autorizadas
`DeleteFile`	`DELETE /v1/files/{id}`

> **Note**
>
> `GetFile`autoriza la descarga de metadatos y contenido. Un director con acceso de solo lectura puede descargar bytes de archivos, no solo archivos de listas.

Habilidades

Action	Rutas 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**
>
> Eliminar una versión de habilidad individual corresponde a`UpdateSkill`, no`DeleteSkill`. Una política que deniegue `aws-external-anthropic:Delete*` aún permite eliminar una versión. `UpdateSkill``CreateSkill`Niégalo y también si necesitas evitar cualquier mutación de habilidad.

Perfiles de usuario

Action	Rutas autorizadas
`CreateUserProfile`	`POST /v1/user_profiles`
`GetUserProfile`	`GET /v1/user_profiles/{id}`
`ListUserProfiles`	`GET /v1/user_profiles`
`UpdateUserProfile`	`POST /v1/user_profiles/{id}`

> **Warning**
>
> La coincidencia de acciones de IAM no distingue entre mayúsculas y minúsculas. El comodín `aws-external-anthropic:*File` coincide con `CreateFileGetFile`, y`DeleteFile`, pero no coincide `ListFiles` (lo que termina en «archivos», no en «archivo»). También coincide en exceso `CreateUserProfileGetUserProfile`, y `UpdateUserProfile` porque «Perfil» termina en «archivo». Si pretendes conceder o denegar únicamente acciones de la API de Files, enuméralas de forma explícita (`CreateFile`,, `GetFileListFiles`,`DeleteFile`) en lugar de utilizar un patrón de `*File` sufijos.

Agentes

Definiciones de agentes para Claude Managed Agents.

Action	Rutas autorizadas
`CreateAgent`	El agente crea rutas
`GetAgent`	El agente lee las rutas
`ListAgents`	Rutas de la lista de agentes
`UpdateAgent`	Rutas de actualización de agentes

Action	Rutas autorizadas
`ArchiveAgent`	Rutas de archivo de agentes

Sesiones

Historial de eventos y sesiones de los agentes.

Action	Rutas autorizadas
`CreateSession`	Sesión: crear rutas
`GetSession`	Rutas de lectura de sesión
`ListSessions`	Rutas de la lista de sesiones
`UpdateSession`	Rutas de actualización de sesiones
`ArchiveSession`	Rutas de archivo de sesiones
`DeleteSession`	Rutas de eliminación de sesiones

Entornos

Configuraciones de contenedores en la nube para sesiones.

Action	Rutas autorizadas
`CreateEnvironment`	Medio ambiente: crea rutas
`GetEnvironment`	El entorno lee las rutas
`ListEnvironments`	Rutas de la lista de entornos
`UpdateEnvironment`	Rutas de actualización del entorno
`ArchiveEnvironment`	Rutas de archivo ambiental

Action	Rutas autorizadas
`DeleteEnvironment`	Entorno: eliminar rutas
`ProcessEnvironmentWork`	Self-hosted rutas de trabajadores del medio ambiente

Almacenes

Bóvedas de credenciales para la autenticación de sesiones.

Action	Rutas autorizadas
`CreateVault`	Crea rutas en Vault
`GetVault`	Vault lee las rutas
`ListVaults`	Lista de rutas de Vault
`UpdateVault`	Rutas de actualización de Vault
`ArchiveVault`	Rutas de archivo de Vault
`DeleteVault`	Vault elimina rutas

> **Note**
>
> Las operaciones de Vault se clasifican como eventos de CloudTrail administración (en lugar de eventos de datos) porque los almacenes contienen credenciales y se benefician del registro de auditoría predeterminado. Otras operaciones de Claude Managed Agents (agentes, sesiones, entornos, almacenes de memoria) son eventos de datos.

Almacenes de memoria

Memoria de agente persistente.

Action	Rutas autorizadas
`CreateMemoryStore`	Almacena memoria para crear rutas
`GetMemoryStore`	Almacena en memoria las rutas de lectura
`ListMemoryStores`	Almacena en memoria las rutas de la lista
`UpdateMemoryStore`	Rutas de actualización del almacén de memoria
`ArchiveMemoryStore`	Rutas de archivo del almacén de memoria
`DeleteMemoryStore`	Almacena en memoria las rutas de eliminación

> **Note**
>
> `GetMemoryStore`lee el contenido de la memoria. Por lo tanto, el `Get*` comodín de una política gestionada o personalizada concede acceso de lectura a la memoria. Limite las políticas de forma explícita si el contenido de la memoria debe restringirse.

Webhooks

Notificaciones de eventos del ciclo de vida de las sesiones.

Action	Rutas 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}`

Action	Rutas autorizadas
`RotateWebhookSecret`	`POST /v1/webhooks/{webhook_endpoint_id}/regenerate_signing_secret`

Espacios de trabajo

Action	Rutas autorizadas
`CreateWorkspace`	`POST /v1/organizations/workspaces`
`GetWorkspace`	`GET /v1/organizations/workspaces/{id}`
`ListWorkspaces`	`GET /v1/organizations/workspaces`
`UpdateWorkspace`	`POST /v1/organizations/workspaces/{id}`
`ArchiveWorkspace`	`POST /v1/organizations/workspaces/{id}/archive`

Al registrarse, se aprovisiona un espacio de trabajo predeterminado; consulte Espacios de trabajo.

Autenticación

Action	Rutas autorizadas
`CallWithBearerToken`	(ninguno)

`CallWithBearerToken`es un permiso de capa de autenticación que autoriza a un principal a autenticarse mediante una clave de API (token de portador) en lugar de mediante AWS SigV4. No se asigna a una ruta. Ofrézcala junto con las acciones del mapa de rutas que deseas que realice el titular de la clave de la API.

Acceso a la consola

Action	Rutas autorizadas
`AssumeConsole`	(ninguno)

`AssumeConsole`autoriza a un director a abrir la consola Claude para una plataforma Claude en el espacio de trabajo de AWS mediante el flujo de federación de consolas de AWS. No se asigna a una ruta. Ofrézcalo a los directores que deberían poder hacer clic en Abrir la consola Claude en la página de servicios de la plataforma Claude en AWS de la consola de AWS. La clave de `aws-external-anthropic:Capability` condición de la `AssumeConsole` acción controla qué capacidad de consola (desarrollador o administrador) puede solicitar un director. Consulte las políticas de IAM para obtener más información y cómo usar la consola Claude para conocer el flujo de inicio de sesión.

Warning

`aws-external-anthropic:AssumeConsole`emite una sesión de Claude Console que dura hasta 12 horas, independientemente de la duración de la sesión de la persona que llama. Una persona que llame y cuya sesión de IAM dure menos de 12 horas puede seguir obteniendo una sesión de consola que dure más tiempo que sus credenciales de origen. Limite este permiso a los directores que necesiten acceder a Claude Console y revoquelo inmediatamente cuando ya no lo necesiten.

Note

Las acciones realizadas dentro de la consola Claude después de la federación no se registran en AWS CloudTrail ni se pueden atribuir a través de IAM. Esto incluye actividades globales relacionadas con el espacio de trabajo, como la visualización de los informes de uso. Si necesita un registro de auditoría de la actividad de la consola, utilice los registros de auditoría disponibles en la consola Claude en lugar de hacerlo. CloudTrail

Route-to-action mapeo

En la siguiente tabla se enumeran todas las rutas de la plataforma Claude en AWS y la acción de IAM necesaria para llamarlas. La acción IAM de la ruta estable también autoriza las solicitudes que utilizan el encabezado. `anthropic-beta` CloudTrail clasifica cada acción como un evento de datos (operaciones de gran volumen en el plano de datos) o como un evento de administración (operaciones del plano de control).

Método	Ruta	Acción de IAM	CloudTrail tipo de evento
`POST`	`/v1/messages`	`CreateInference`	Datos
`POST`	`/v1/messages/count_tokens`	`CountTokens`	Datos
`POST`	`/v1/messages/batches`	`CreateBatchInference`	Datos
`GET`	`/v1/messages/batches`	`ListBatchInferences`	Datos
`GET`	`/v1/messages/batches/{id}`	`GetBatchInference`	Datos
`GET`	`/v1/messages/batches/{id}/results`	`GetBatchInference`	Datos
`POST`	`/v1/messages/batches/{id}/cancel`	`CancelBatchInference`	Datos
`DELETE`	`/v1/messages/batches/{id}`	`DeleteBatchInference`	Datos
`GET`	`/v1/models`	`ListModels`	Datos
`GET`	`/v1/models/{id}`	`GetModel`	Datos
`POST`	`/v1/files`	`CreateFile`	Datos

Método	Ruta	Acción de IAM	CloudTrail tipo de evento
GET	/v1/files	ListFiles	Datos
GET	/v1/files/{id}	GetFile	Datos
GET	/v1/files/{id}/content	GetFile	Datos
DELETE	/v1/files/{id}	DeleteFile	Datos
POST	/v1/skills	CreateSkill	Datos
GET	/v1/skills	ListSkills	Datos
GET	/v1/skills/{id}	GetSkill	Datos
DELETE	/v1/skills/{id}	DeleteSkill	Datos
POST	/v1/skills/{id}/versions	UpdateSkill	Datos
GET	/v1/skills/{id}/versions	GetSkill	Datos
GET	/v1/skills/{id}/versions/{version}	GetSkill	Datos
GET	/v1/skills/{id}/versions/{version}/content	GetSkill	Datos
DELETE	/v1/skills/{id}/versions/{version}	UpdateSkill	Datos
POST	/v1/user_profiles	CreateUserProfile	Datos
GET	/v1/user_profiles	ListUserProfiles	Datos
GET	/v1/user_profiles/{id}	GetUserProfile	Datos
POST	/v1/user_profiles/{id}	UpdateUserProfile	Datos

Método	Ruta	Acción de IAM	CloudTrail tipo de evento
`POST`	`/v1/organizations/workspaces`	`CreateWorkspace`	Administración
`GET`	`/v1/organizations/workspaces`	`ListWorkspaces`	Administración
`GET`	`/v1/organizations/workspaces/{id}`	`GetWorkspace`	Administración
`POST`	`/v1/organizations/workspaces/{id}`	`UpdateWorkspace`	Administración
`POST`	`/v1/organizations/workspaces/{id}/archive`	`ArchiveWorkspace`	Administración

Las rutas de Claude Managed Agents (agentes, sesiones, entornos, bóvedas, almacenes de memoria) siguen el mismo patrón de ruta a la acción; para obtener un mapeo completo por ruta, consulte la referencia sobre las acciones de Anthropic IAM. Las rutas de Vault son eventos de administración; las rutas de agente, sesión, entorno y almacenamiento de memoria son eventos de datos.

De forma predeterminada, las rutas que no se encuentran en la plataforma Claude en AWS se deniegan en la puerta de enlace.

Véase también

- Políticas de IAM (por ejemplo, políticas y políticas administradas)
- Guía del usuario de AWS IAM para la sintaxis de las políticas de IAM y la lógica de evaluación
- Guía CloudTrail del usuario de AWS para la configuración de registros de auditoría

Historial de revisión

En la siguiente tabla se describen los cambios importantes en la guía del usuario de la plataforma Claude en AWS.

Cambio	Descripción	Fecha
Publicación inicial	Publicó la guía del usuario de la plataforma Claude en AWS.	7 de mayo de 2026

Las traducciones son generadas a través de traducción automática. En caso de conflicto entre la traducción y la version original de inglés, prevalecerá la version en inglés.