aws

Panduan Pengguna

Platform Claude di AWS

Platform Claude di AWS: Panduan Pengguna

Table of Contents

Apa itu Platform Claude di AWS?

Akses kapabilitas platform lengkap Claude melalui AWS dengan Anthropic-managed infrastruktur.

Platform Claude di AWS memberi Anda pengalaman platform Antropik lengkap, termasuk API Pesan, Keterampilan Agen, eksekusi kode, dan fitur beta, yang dapat diakses melalui akun AWS Anda. Tidak seperti Amazon Bedrock, di mana AWS mengoperasikan tumpukan inferensi, Anthropic mengoperasikan Platform Claude di AWS. AWS menyediakan lapisan otentikasi (SIgV4 atau kunci API), kontrol IAM-based akses, dan integrasi penagihan melalui AWS Marketplace.

> **Note**
>
> Anthropic SDK mendukung Platform Claude di AWS. Untuk ketersediaan klien per bahasa, lihat dokumentasi SDK Klien Antropik.

Cara kerja integrasi platform

Model Claude berjalan pada Anthropic-managed infrastruktur. Ini adalah integrasi komersial untuk penagihan dan akses melalui AWS. Baik Anthropic dan AWS bertindak sebagai pemroses data independen. Ketentuan Layanan AWS mengatur Platform Claude di AWS. Ketentuan Layanan Komersil Anthropic, Adendum Pemrosesan Data, dan Kebijakan Penggunaan juga berlaku.

Perhatikan karakteristik operasional berikut. Data mungkin tidak berada di AWS. Inferensi dapat merutekan ke awan utama Anthropic. Sublayanan dapat berubah tanpa pemberitahuan. Tetapkan `inference_geo` parameter per permintaan untuk menyematkan inferensi ke geografi tertentu. Lihat Data residensi untuk detailnya.

Platform Claude di AWS mengikuti kebijakan retensi data yang sama dengan Claude API pihak pertama. Zero Data Retention (ZDR) tersedia berdasarkan permintaan. Hubungi perwakilan akun Anthropic Anda untuk mengaktifkannya untuk akun Anda.

Fitur Platform Claude di AWS

Platform Claude di AWS menyediakan kemampuan berikut:

- Same-day model dan akses fitur - Model Claude baru dan fitur API tersedia pada hari yang sama saat diluncurkan di Claude API pihak pertama.

- Keterampilan Agen — Gunakan Keterampilan bawaan untuk pembuatan dokumen (PowerPoint, Excel, Word, PDF) dan buat Keterampilan khusus.
- Eksekusi kode - Jalankan kode di kotak pasir terkelola Anthropic.
- Pemikiran yang diperluas — Aktifkan pemikiran yang diperluas untuk tugas penalaran yang kompleks.
- Streaming dan pemrosesan batch - Gunakan streaming SSE waktu nyata atau kirimkan permintaan batch untuk beban kerja throughput tinggi.
- Prompt caching — Alat cache, prompt sistem, dan riwayat pesan untuk mengurangi latensi dan biaya.
- Files API — Unggah dan referensi file di seluruh permintaan.
- Integrasi AWS IAM — Kontrol akses dengan kebijakan IAM standar dan otentikasi SiGv4.
- Penagihan AWS Terpadu — Bayar melalui akun AWS Anda yang ada dengan pengukuran berbasis konsumsi (Marketplace sebagai backend penagihan).

Untuk daftar lengkap fitur yang didukung, lihat Dukungan fitur.

Bagaimana panduan ini diatur

Panduan ini mencakup topik-topik berikut:

- Memulai — Penyiapan akun, prasyarat, otentikasi, dan instalasi SDK.
- Menggunakan API - Model yang tersedia, membuat permintaan, dan dukungan fitur.
- Mengelola lingkungan Anda — Residensi data, ruang kerja, Konsol Claude, batas tarif, dan penagihan.
- Pemantauan dan operasi — CloudTrail pencatatan dan permintaan pelacakan ID.
- Migrasi — Pindah dari Amazon Bedrock ke Claude Platform di AWS.
- Keamanan dan kontrol akses — referensi kebijakan dan tindakan IAM.

Platform Claude di AWS vs Amazon Bedrock

Kedua penawaran memungkinkan Anda menggunakan Claude melalui AWS, tetapi keduanya berbeda secara signifikan dalam arsitektur, permukaan API, dan ketersediaan fitur.

	Platform Claude di AWS	Amazon Bedrock
Siapa yang mengoperasikan tumpukan	Antropik	AWS
Permukaan API	API Pesan Antropik () `/v1/messages`	API Batuan Dasar (Converse/) InvokeModel
Ketersediaan fitur	Same-day sebagai Claude API	Tergantung pada timeline integrasi AWS
Keterampilan Agen	Available	Tidak tersedia (memerlukan eksekusi kode)
Fitur beta	Lewati dengan `anthropic-beta` header (lihat Dukungan fitur)	Membutuhkan dukungan AWS
Autentikasi	AWS IAM/SIgV4 atau kunci API	AWS IAM/SiGv4 atau token pembawa (hanya SDK C #, Go, dan Java)
URL dasar	`aws-external-anthropic.{region}.api.aws`	`bedrock-runtime.{region}.amazonaws.com`
Klien SDK	Platform-specific kelas klien (misalnya, `AnthropicAWS` dalam Python), dalam versi beta	`AnthropicBedrock` /SDK Batuan Dasar
Konsol	Claude Console (`platform.claude.com` , akses melalui AWS Console)	Konsol Batuan Dasar

	Platform Claude di AWS	Amazon Bedrock
Batas nilai dan kuota	Dikelola oleh Anthropic	Dikelola oleh AWS
Pemroses data	Antropik dan AWS	AWS

Jika Anda membutuhkan AWS-operated Claude dengan API Bedrock, lihat Amazon Bedrock.

> **Important**
>
> Anthropic menyediakan Platform Claude di AWS sebagai penawaran pihak ketiga. Ini tidak tercakup dalam program kepatuhan AWS standar, sertifikasi, atau laporan audit (seperti kelayakan SOC, ISO, atau HIPAA). Pelanggan bertanggung jawab penuh untuk melakukan uji tuntas mereka sendiri untuk memastikan bahwa penawaran pihak ketiga ini memenuhi persyaratan peraturan, hukum, dan kepatuhan mereka.

Kapan memilih Bedrock: Pilih Amazon Bedrock jika organisasi Anda memerlukan AWS-operated inferensi, AWS sebagai pemroses data tunggal, atau cakupan di bawah program kepatuhan AWS (termasuk kelayakan FedRAMP High, IL4, IL5, SOC, ISO, dan HIPAA). Bedrock berjalan sepenuhnya pada AWS-controlled infrastruktur dengan AWS sebagai pihak operasi.

Siapkan akun Anda

Menyiapkan Platform Claude di AWS memiliki empat fase: mendaftar di AWS Console, mengatur organisasi Anthropic Anda, mencatat ID ruang kerja Anda, dan masuk ke Konsol Claude.

> **Note**
>
> Mendaftar melalui AWS Console menyediakan organisasi Anthropic baru yang terkait dengan akun AWS Anda. Organisasi ini terpisah dari organisasi yang ada yang dimiliki perusahaan Anda dengan Anthropic, termasuk organisasi Claude Enterprise yang diperoleh melalui AWS Marketplace. Kunci API, ruang kerja, dan setelan Konsol Claude dari organisasi Antropik pihak pertama tidak terbawa.
> Jika Anda memiliki penawaran pribadi Amazon Bedrock yang sudah ada, hubungi eksekutif akun Anthropic atau AWS Anda sebelum mendaftar sehingga diskon Anda berlaku dari permintaan pertama Anda. Diskon tidak dapat diterapkan secara surut untuk penggunaan yang terjadi sebelum penawaran pribadi Anda diterima. Lihat Penawaran pribadi.

Langkah 1: Daftar di AWS Console

1. Buka AWS Console dan navigasikan ke halaman layanan Platform Claude di AWS.
2. Pilih Daftar.
3. Pilih Lanjutkan.

Halaman ini menampilkan spanduk Sign-up yang sedang berlangsung. Tetap di halaman. Sign-up membutuhkan waktu beberapa menit sementara AWS menangani langganan AWS Marketplace untuk Anda, lalu mengarahkan Anda secara otomatis.

Jika organisasi Anda memiliki penawaran pribadi dari Anthropic, Konsol akan mencarinya dan meminta Anda untuk menerimanya di AWS Marketplace. Lihat Penawaran pribadi untuk detailnya.

> **Note**
>
> Jika Anda menggunakan Platform Claude di AWS, konten Anda (seperti permintaan dan penyelesaian) diproses oleh Anthropic di luar AWS. Lihat kebijakan penggunaan data Anthropic untuk detail tentang cara konten dan metadata diproses dan disimpan.

Langkah 2: Siapkan organisasi Antropik Anda

Setelah pendaftaran selesai, Anda diarahkan ke. `platform.claude.com/partner-signup`

1. Masukkan alamat email pemilik organisasi Anda dan pilih Memulai.
2. Periksa kotak masuk email tersebut untuk tautan pengaturan dan ikuti. Jika browser Anda menampilkan halaman Masuk sebagai akun lain, pilih Keluar dan lanjutkan.
3. Lengkapi formulir detail organisasi (nama organisasi, jenis entitas, negara, tujuan penggunaan) dan pilih Penyiapan lengkap.

Menyelesaikan penyiapan menciptakan organisasi Antropik Anda. Ketentuan Layanan AWS mengatur Platform Claude di AWS. Ketentuan Layanan Komersil Anthropic, Adendum Pemrosesan Data, dan Kebijakan Penggunaan juga berlaku. Halaman layanan AWS Console sekarang menampilkan navigasi kiri dengan Home, kunci API, Quickstart, dan Workspaces.

Langkah 3: Perhatikan ID ruang kerja Anda

Ruang kerja default disediakan secara otomatis saat Anda mendaftar. Ruang kerja tambahan dapat dibuat per wilayah; lihat Ruang kerja untuk detail tentang pengikatan wilayah, manajemen ruang kerja, dan pelingkupan sumber daya IAM.

Temukan ID ruang kerja di bawah Workspaces di AWS Console Claude Platform pada halaman layanan AWS atau di Konsol Claude (lihat Menggunakan Konsol Claude). ID ruang kerja menggunakan format yang `wrkspc_` diikuti oleh pengenal alfanumerik.

Langkah 4: Masuk ke Konsol Claude

Akses ke Konsol Claude digabungkan melalui AWS IAM:

1. Asumsikan peran IAM dengan `aws-external-anthropic:AssumeConsole` izin. Lihat tindakan IAM.
2. Dari halaman layanan Platform Claude di AWS, pilih Masuk. AWS Console mengeluarkan JWT dan mengarahkan Anda ke. `platform.claude.com`
3. Saat masuk pertama, Anda akan diminta untuk mendapatkan alamat email. Masukkan email kerja Anda. Platform menyediakan pengguna Claude Console Anda tepat waktu.

Saat Anda masuk melalui AWS Console, Konsol Claude akan tercakup ke Platform Claude Anda di organisasi AWS. Indikator Akun yang dikelola oleh AWS muncul di bilah sisi Claude Console.

Memecahkan masalah pengaturan akun

- "Sign-up gagal: Gagal mengaktifkan OutboundWebIdentityFederation “: Jika Anda melihat spanduk merah ini pada kiriman pertama, pilih Lanjutkan lagi. Pemberdayaan IAM dapat memakan waktu sejenak untuk menyebar.
- Tidak ada indikator kemajuan selama pendaftaran: Sign-up membutuhkan waktu beberapa menit. Halaman menampilkan spanduk statis Sign-up yang sedang berlangsung tanpa bilah kemajuan sementara AWS menyediakan akun Anda.
- “Masuk sebagai akun yang berbeda” setelah mengikuti tautan penyiapan: Pilih Keluar dan lanjutkan. Halaman tersebut mengautentikasi ulang Anda dengan alamat email yang Anda masukkan.
- Pesan “Tidak ditemukan” saat masuk: Pesan ini mungkin muncul sebentar selama pengalihan. Anda dapat mengabaikannya.
- Halaman penggunaan tidak menampilkan data setelah panggilan API pertama Anda: Data penggunaan dapat memakan waktu beberapa menit untuk muncul di Konsol Claude.

Prasyarat

Sebelum melakukan panggilan API, pastikan Anda memiliki:

1. Akun AWS aktif dengan berlangganan Platform Claude di AWS (lihat Mengatur akun Anda).
2. AWS CLI diinstal dan dikonfigurasi.
3. Federasi identitas web keluar diaktifkan di akun AWS Anda (langkah penyiapan satu kali; lihat Mengaktifkan federasi identitas web keluar).
4. ID ruang kerja Anda (lihat Mendapatkan ID ruang kerja Anda).

Aktifkan federasi identitas web keluar

Platform Claude di gateway AWS memanggil `sts:GetWebIdentityToken` sisi server untuk mencetak JWT yang diteruskan ke Anthropic. Kemampuan STS ini dinonaktifkan secara default di setiap akun AWS. Aktifkan sekali per akun:

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

Jika responsnya`[ERROR] (FeatureEnabled) … already enabled` , pengaturan sudah aktif untuk akun Anda dan Anda dapat melanjutkan. Verifikasi dan ambil URL penerbit akun Anda:

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

> **Warning**
>
> Tanpa langkah ini, setiap permintaan kembali`"Outbound web identity federation is disabled for your account"`. Ini adalah kesalahan pengaturan yang paling umum.

Dapatkan ID ruang kerja Anda

Ruang kerja default disediakan secara otomatis di setiap wilayah saat Anda mendaftar (lihat Mengatur akun). Ruang kerja terikat pada satu wilayah AWS. Anda dapat menemukan ID ruang kerja di Claude Console di bawah Workspaces, atau di bagian Workspaces pada halaman layanan AWS Console. Lihat Ruang kerja untuk pengikatan wilayah dan pelingkupan sumber daya IAM.

Atur variabel `ANTHROPIC_AWS_WORKSPACE_ID` dan `AWS_REGION` lingkungan sehingga klien SDK membacanya secara otomatis:

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

Wilayah ini diperlukan. Klien SDK melempar jika tidak ada wilayah yang disetel. Lulus`aws_region`/`awsRegion`ke konstruktor, atau set `AWS_REGION` (atau`AWS_DEFAULT_REGION`). Semua wilayah komersial AWS didukung; wilayah keikutsertaan mengharuskan Anda memilih akun terlebih dahulu ke wilayah tersebut.

Autentikasi

Platform Claude di AWS mendukung dua metode otentikasi: AWS IAM dengan penandatanganan permintaan SigV4 (primer) dan otentikasi kunci API. Keduanya menggunakan URL dasar dan format permintaan yang sama.

SDK Antropik (lihat Menginstal SDK) mengimplementasikan alur otentikasi dan resolusi kredensyal yang dijelaskan di bawah ini. Jika Anda tidak menggunakan Anthropic SDK, Anda harus membuat SigV4-signed permintaan (atau menampilkan kunci API Anda sebagai token pembawa) terhadap titik akhir regional. `https://aws-external-anthropic.<region>.api.aws` Untuk konfigurasi SDK-specific klien dan urutan resolusi kredensyal otoritatif, lihat panduan penyiapan Claude on AWS di dokumentasi Anthropic.

Otentikasi SiGv4

SiGv4 adalah jalur asli perusahaan dan terintegrasi dengan kebijakan, peran, dan audit AWS IAM Anda yang ada. Konfigurasikan kredensyal AWS menggunakan metode apa pun yang didukung oleh rantai penyedia kredensi default AWS:

- Variabel lingkungan (`AWS_ACCESS_KEY_ID`,`AWS_SECRET_ACCESS_KEY`,`AWS_SESSION_TOKEN`)
- File kredensyal bersama () `~/.aws/credentials`
- File konfigurasi bersama (`~/.aws/config`) termasuk SSO dan `credential_process`
- Identitas web (`AWS_WEB_IDENTITY_TOKEN_FILE`dan`AWS_ROLE_ARN`) untuk IRSA dan Tindakan GitHub
- Kredensyal kontainer ECS
- Layanan metadata instans EC2 (IMDS)

Untuk memverifikasi bahwa Anthropic SDK mengambil kredensyal Anda dan menyelesaikan ke titik akhir regional yang benar, buat permintaan pengujian dengan mengikuti contoh dalam Membuat permintaan. Respons yang berhasil mengonfirmasi bahwa kredensyal, wilayah, URL dasar, dan ID ruang kerja semuanya dikonfigurasi dengan benar.

Otentikasi kunci API

Untuk jalur integrasi yang lebih sederhana (pengembangan dan skrip lokal), Anda dapat mengautentikasi dengan kunci API alih-alih SigV4. Setel variabel `ANTHROPIC_AWS_API_KEY`

lingkungan atau teruskan `apiKey` ke konstruktor SDK. Anthropic SDK mengirimkan kunci sebagai token pembawa ke Platform Claude pada titik akhir AWS; IAM mengotorisasi permintaan melalui `aws-external-anthropic:CallWithBearerToken` tindakan (lihat kebijakan IAM).

Buat kunci API di Konsol AWS di bawah Platform Claude di AWS → kunci API. Pilih Hasilkan kunci, lalu salin nilai kuncinya. Berikan tindakan `aws-external-anthropic:CallWithBearerToken` IAM kepada prinsipal yang harus diizinkan untuk menggunakan otentikasi kunci API.

> **Note**
>
> Hanya kunci API yang dibuat di Konsol AWS di bawah Platform Claude di AWS yang berfungsi dengan layanan ini.
>
> - Kunci yang dibuat di Konsol Claude standar untuk akses API pihak pertama tidak berfungsi terhadap Platform Claude di titik akhir AWS.
> - Kunci Amazon Bedrock API juga tidak berfungsi — Bedrock menggunakan endpoint terpisah, alur autentikasi, dan namespace IAM.

Short-term Kunci API

Untuk kasus di mana Anda menginginkan otentikasi kunci API tetapi tanpa masa pakai kunci yang berumur panjang, Anthropic menerbitkan pustaka generator token yang mencetak kunci API jangka pendek dari kredensyal AWS IAM Anda. Token default untuk masa pakai 12 jam dan dapat dicakup ke ruang kerja tertentu dan tindakan. `aws-external-anthropic:CallWithBearerToken` Instal generator untuk bahasa Anda, tukarkan kredensyal IAM dengan kunci API, dan teruskan kunci ke Anthropic SDK.

- Python: -generator-untuk-aws-eksternal-anthropic-python aws/token
- JavaScript / TypeScript: aws/token -generator-untuk-aws-eksternal-antropik-js
- Java: aws/token-generator-untuk-aws-eksternal-antropik-java

Short-term Kunci API masih memerlukan prinsipal IAM pemanggil untuk bertahan `aws-external-anthropic:CallWithBearerToken` di ruang kerja target. Mereka kedaluwarsa sendiri dan tidak perlu diputar atau dicabut secara eksplisit.

Prioritas kredensi dan resolusi wilayah

Platform Claude Anthropic SDK pada klien AWS menyelesaikan kredensyal dan wilayah menggunakan urutan prioritas yang ditentukan. Nama argumen mengikuti konvensi masing-masing bahasa: camelCase for dan PHP, snake_case TypeScript untuk Python dan Ruby, untuk Go PascalCase , dan properti atau pola pembangun untuk C# dan Java.

Untuk urutan prioritas otoritatif, variabel lingkungan yang didukung, dan argumen konstruktor untuk setiap SDK, lihat Claude on AWS setup dalam dokumentasi Anthropic. Secara umum, argumen konstruktor eksplisit lebih diutamakan daripada variabel lingkungan, dan lebih diutamakan daripada rantai `ANTHROPIC_AWS_API_KEY` penyedia kredensi AWS default. Wilayah diperlukan; tidak seperti `AnthropicBedrock` (yang kembali ke`us-east-1`), klien Claude on AWS melempar jika tidak ada wilayah yang disediakan oleh konstruktor atau oleh/. `AWS_REGION` `AWS_DEFAULT_REGION`

ID Ruang Kerja

Setiap permintaan bidang data harus menyertakan ID ruang kerja di `anthropic-workspace-id` header. Anthropic SDK membaca ini dari variabel `ANTHROPIC_AWS_WORKSPACE_ID` lingkungan secara default, atau Anda dapat meneruskan`workspaceId`/`workspace_id`ke konstruktor klien. Jika Anda menggunakan `Anthropic` klien dasar (bukan Claude-on-AWS klien), berikan header secara eksplisit. Lihat Membuat permintaan untuk contoh per bahasa dan Ruang Kerja untuk mengetahui cara menemukan ID ruang kerja Anda.

Instal SDK

SDK klien Anthropic mendukung Platform Claude di AWS. Ketujuh SDK menyediakan kelas klien khusus platform; sebagai alternatif, Anda dapat mengonfigurasi klien dasar dengan Platform Claude pada URL dasar AWS.

Python

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

> **Tip**
>
> Di macOS dengan Homebrew Python atau lingkungan `pip install` Python yang dikelola secara eksternal lainnya, mungkin gagal dengan kesalahan PEP 668. `externally-managed-environment` Buat dan aktifkan lingkungan virtual terlebih dahulu:`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**
>
> Klien SDK untuk Claude Platform di AWS dalam versi beta.

Model yang tersedia

Platform Claude di AWS menawarkan set model Claude yang sama dengan Claude API pihak pertama.

Untuk daftar model, ID model, ukuran jendela konteks, dan kemampuan saat ini, lihat Ikhtisar model dalam dokumentasi Antropik.

ID Model

ID model pada Platform Claude di AWS identik dengan yang digunakan pada Claude API pihak pertama (misalnya,,,). `claude-opus-4-6` `claude-sonnet-4-6` `claude-haiku-4-5` Tidak ada Bedrock-style ARN atau `anthropic.` awalan — gunakan ID model persis seperti Anthropic menerbitkannya.

Membuat permintaan

Platform Claude di AWS menggunakan Anthropic Messages API (`/v1/messages`), permukaan API yang sama dengan platform pihak pertama Claude. Perbedaannya adalah URL dasar, metode otentikasi, dan `anthropic-workspace-id` header wajib yang mengidentifikasi ruang kerja mana yang ditargetkan permintaan.

Cangkang

```
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
```

Klien membaca `AWS_REGION` (atau`AWS_DEFAULT_REGION`) dan `ANTHROPIC_AWS_WORKSPACE_ID` dari lingkungan. Anda dapat mengganti baik dengan meneruskan`aws_region`/`awsRegion`atau`workspace_id`/`workspaceId`ke konstruktor. ID wilayah dan ruang kerja diperlukan; konstruktor melempar jika salah satu tidak dapat diselesaikan dari sumber-sumber ini.

> **Note**
>
> `x-amz-security-token`Header (curl) hanya diperlukan untuk kredensi sementara seperti peran IAM, SSO, atau STS. Hilangkan saat menggunakan kredensil pengguna IAM jangka panjang. Klien SDK menangani ini secara otomatis berdasarkan sumber kredensialnya.

`--aws-sigv4`Nilai mengikuti format`aws:amz:<region>:<service>`. Nama layanan SigV4 adalah`aws-external-anthropic`, dan wilayah harus cocok dengan wilayah di URL titik akhir Anda. Ketidakcocokan di keduanya menghasilkan kesalahan penolakan tanda tangan generik daripada diagnostik tertentu.

Menggunakan klien Antropik dasar

Jika Anda memilih untuk tidak menambahkan dependensi AWS SDK, Anda dapat menggunakan klien dasar`Anthropic`. Jalur ini menggunakan nama variabel lingkungan SDK vanilla daripada `ANTHROPIC_AWS_*` nama yang dibaca klien khusus:

```
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'
```

Python, TypeScript, dan Go SDK membaca `ANTHROPIC_API_KEY` dan `ANTHROPIC_BASE_URL` secara otomatis; untuk bahasa lain, teruskan ke konstruktor. Dalam setiap bahasa, berikan ID ruang kerja di `anthropic-workspace-id` header.

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);
```

Dukungan fitur

Platform Claude di AWS menggunakan Anthropic Messages API secara langsung. Anda mendapatkan paritas fitur lengkap dengan Claude API pihak pertama, kecuali jika disebutkan dalam Fitur yang saat ini tidak tersedia:

- Fitur beta: Lewati `anthropic-beta` header standar untuk mengakses fitur beta, seperti yang Anda lakukan dengan Claude API.
- Keterampilan Agen: Gunakan Keterampilan Agen bawaan dan kustom dengan `container.skills` parameter dan header beta yang sama dengan Claude API. Semua Keterampilan pra-bangun (PowerPoint, Excel, Word, PDF) bekerja di luar kotak.
- Eksekusi kode: Jalankan kode di kotak pasir terkelola Anthropic menggunakan alat eksekusi kode.
- Penggunaan alat: Penggunaan komputer dan semua kemampuan penggunaan alat lainnya tersedia.
- Pemikiran yang diperluas: Aktifkan pemikiran yang diperluas dengan parameter yang sama dengan Claude API.
- Streaming: Dukungan streaming SSE penuh untuk respons waktu nyata.
- Pemrosesan batch: Kirim permintaan batch untuk beban kerja throughput tinggi.
- Prompt caching: Alat cache, prompt sistem, dan riwayat pesan untuk mengurangi latensi dan biaya. Semua kemampuan caching cepat (TTL 5 menit, TTL 1 jam, dan caching otomatis) tersedia.
- File API: Unggah dan referensi file di seluruh permintaan.
- Tag ruang kerja dengan integrasi IAM: Ruang kerja dapat diberi tag, dan tag mengalir ke IAM untuk kontrol akses berbasis atribut dan ke Laporan Biaya dan Penggunaan AWS untuk alokasi biaya. Penandaan ruang kerja adalah fitur GA di Platform Claude di AWS (ini adalah fitur beta pada Claude API pihak pertama).

Agen Dikelola Claude

Agen Terkelola Claude tersedia di Platform Claude di AWS. Agen, sesi, lingkungan, brankas kredensi, dan penyimpanan memori adalah sumber daya IAM kelas satu. Lihat Tindakan IAM untuk referensi tindakan dan Menggunakan Konsol Claude untuk halaman terkait. Anthropic Agent SDK bekerja dengan Claude Platform di AWS tanpa integrasi tambahan. Arahkan ke Platform Claude pada URL dasar AWS dan autentikasi dengan SigV4 atau kunci API.

Sesi otonom di Platform Claude di AWS memerlukan autentikasi ulang setiap 6 jam. Long-running agen yang berjalan harus menyegarkan kredensi SigV4 atau kunci API mereka di dalam jendela itu, atau sesi berakhir.

Kemampuan Agen Terkelola Claude berikut saat ini tidak tersedia di Platform Claude di AWS:

- Hasil: Pelacakan hasil untuk sesi agen tidak tersedia.
- Multi-agent sesi: Sesi dengan beberapa agen yang berinteraksi tidak tersedia.

Kepatuhan

> **Important**
>
> Anthropic menyediakan layanan ini sebagai penawaran pihak ketiga. Ini tidak tercakup dalam program kepatuhan AWS standar, sertifikasi, atau laporan audit (seperti kelayakan SOC, ISO, atau HIPAA). Pelanggan bertanggung jawab penuh untuk melakukan uji tuntas mereka sendiri untuk memastikan bahwa penawaran pihak ketiga ini memenuhi persyaratan peraturan, hukum, dan kepatuhan mereka. Sebelum memproses data sensitif atau teregulasi, Anda harus meninjau dokumentasi kepatuhan resmi Anthropic melalui Anthropic Trust Center. Lihat Ketentuan Layanan AWS untuk informasi selengkapnya.

Bagaimana keanggotaan ruang kerja dimodelkan

Pada Claude API pihak pertama, akses diatur oleh keanggotaan pengguna-ke-ruang kerja — pengguna ditambahkan ke ruang kerja dan mewarisi hak akses ruang kerja. Platform Claude di AWS menggantikan model tersebut dengan otorisasi IAM: tidak ada daftar pengguna per ruang kerja. Sebagai gantinya, kepala sekolah IAM (pengguna atau peran) memegang kebijakan yang memberikan `aws-external-anthropic` tindakan terhadap ARN ruang kerja tertentu. Evaluasi kebijakan IAM menentukan apa yang dapat dilakukan setiap kepala sekolah di setiap ruang kerja.

Berikan dan cabut akses ruang kerja dengan memodifikasi kebijakan IAM (SCP, batas izin, kebijakan yang dilampirkan identitas, atau kondisi tingkat sumber daya) daripada mengelola keanggotaan per ruang kerja di Konsol Claude. Lihat kebijakan IAM untuk contoh.

Fitur saat ini tidak tersedia

Kemampuan berikut saat ini tidak tersedia di Platform Claude di AWS:

- Kesiapan HIPAA: HIPAA-ready Program Anthropic tidak tersedia di Platform Claude di AWS. Pelanggan dengan persyaratan HIPAA harus mengevaluasi Claude di Amazon Bedrock sebagai gantinya.
- Tingkat layanan di luar Standar dan Batch: Tingkat Prioritas tidak tersedia.
- Admin API — anggota organisasi, anggota ruang kerja, undangan, kunci API, laporan penggunaan, laporan biaya, dan titik akhir laporan batas tarif: Titik akhir API Admin ini saat ini tidak tersedia. Workspace CRUD dan rename endpoints (`/v1/organizations/workspaces`) tersedia melalui `aws-external-anthropic` namespace; lihat Tindakan IAM. Keanggotaan dimodelkan melalui AWS IAM daripada melalui daftar keanggotaan ruang kerja (lihat bagian sebelumnya). Siklus hidup kunci API dikelola di Konsol AWS di bawah Platform Claude di AWS → kunci API. Untuk data batas penggunaan, biaya, dan tarif, gunakan tampilan Claude Console (lihat Pemantauan dan pencatatan) atau Anthropic Usage and Cost API.
- Batas pengeluaran: Tidak tersedia. Andalkan kontrol penagihan AWS sebagai gantinya.
- Ruang kerja Claude Code dan API Analytics: Ruang kerja Claude Code dengan batas tarif otomatis tidak tersedia. Penggunaan Kode Claude muncul di tampilan penggunaan umum daripada layar khusus.
- Otentikasi OAuth: Tidak didukung. Gunakan SiGv4 atau otentikasi kunci API.
- OpenAI-compatible Titik akhir API: Tidak tersedia di Platform Claude di AWS.
- Workspace-level kontrol geografi inferensi: `allowed_inference_geos` dan tidak `default_inference_geo` tersedia. Tetapkan `inference_geo` pada setiap permintaan sebagai gantinya.

Data residensi

Platform Claude di AWS mendukung geografi inferensi berikut:

- AS: Inferensi tetap berada di dalam pusat data AS. Pengganda harga 1.1x berlaku.
- Global: Inferensi dapat merutekan ke pusat Anthropic-operated data mana pun di seluruh dunia. Harga standar berlaku.

Atur geografi inferensi per permintaan dengan parameter: `inference_geo`

> **Note**
>
> `inference_geo`Parameter ini didukung pada Claude Opus 4.6, Claude Sonnet 4.6, dan model yang lebih baru. Permintaan dengan `inference_geo` Claude Opus 4.5, Claude Sonnet 4.5, atau Claude Haiku 4.5 mengembalikan kesalahan 400. Lihat Residensi data untuk detail ketersediaan model.

Cangkang

```
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
```

Jika Anda menghilangkan`inference_geo`, permintaan default ke. `global`

Workspace-level Kontrol geografi inferensi (`allowed_inference_geos`dan`default_inference_geo`) tidak tersedia di Platform Claude di AWS. Tetapkan `inference_geo` pada setiap permintaan sebagai gantinya.

Workspace

Inferensi dan permintaan sumber daya di Platform Claude di AWS menargetkan ruang kerja. Anda meneruskan ID ruang kerja di `anthropic-workspace-id` header pada panggilan API ini. ID ruang kerja menggunakan format yang ditandai `wrkspc_` diikuti oleh pengidentifikasi alfanumerik (misalnya,). `wrkspc_01AbCdEf23GhIj` Lihat Mendapatkan ID ruang kerja Anda jika Anda belum memilikinya.

Pelingkupan ruang kerja

Ruang kerja terikat pada satu wilayah AWS. Ruang kerja yang dibuat hanya `us-west-2` dapat diakses melalui titik `us-west-2` akhir. Penggunaan, kuota, biaya, file, batch, dan Keterampilan semuanya digulung per ruang kerja, memberi Anda rincian per wilayah di Konsol Claude.

Ruang kerja juga berfungsi sebagai sumber daya IAM utama untuk Platform Claude di AWS. Anda memberikan atau menolak akses ke ruang kerja tertentu melalui kebijakan AWS IAM menggunakan ARN ruang kerja. Segmen sumber daya ARN adalah ID `wrkspc_` -prefixed yang sama yang Anda berikan di header: `anthropic-workspace-id`

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

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

Lihat kebijakan IAM untuk contoh kebijakan.

Membuat ruang kerja

Ruang kerja default disediakan secara otomatis saat mendaftar. Ruang kerja tambahan dapat dibuat, diperbarui, diganti namanya, dan diarsipkan melalui `/v1/organizations/workspaces` titik akhir, yang diizinkan oleh tindakan yang sesuai (, `CreateWorkspaceUpdateWorkspace,ArchiveWorkspace`); lihat `aws-external-anthropic` tindakan IAM.

Mengatur ID ruang kerja di klien Anda

Setiap permintaan bidang data harus menyertakan ID ruang kerja di `anthropic-workspace-id` header. Saat Anda menggunakan Claude-on-AWS klien Anthropic SDK, ada tiga cara untuk menyediakannya:

1. Variabel lingkungan - set`ANTHROPIC_AWS_WORKSPACE_ID`, dan klien membacanya secara otomatis.

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

2. Argumen konstruktor — pass`workspaceId`/`workspace_id`(nama mengikuti konvensi bahasa SDK) saat membuat instance klien.
3. Per-request override — teruskan header secara langsung pada permintaan individual jika Anda perlu menangani beberapa ruang kerja dari klien yang sama.

Jika Anda menggunakan `Anthropic` klien dasar (tanpa backend AWS), setel `anthropic-workspace-id` header secara eksplisit pada setiap permintaan — lihat Membuat permintaan untuk contoh per bahasa. Untuk nama SDK-specific argumen, lihat dokumentasi Anthropic Client SDK.

Menandai ruang kerja

Tag ruang kerja adalah pasangan nilai kunci yang dilampirkan ke ruang kerja. Mereka terintegrasi dengan AWS IAM untuk kontrol akses berbasis atribut, dan dengan Laporan Biaya dan Penggunaan AWS untuk alokasi biaya (lihat Pemantauan dan pencatatan untuk detail CUR). Penandaan adalah GA di Platform Claude di AWS.

Perbarui tag pada ruang kerja yang ada dengan `TagResource` dan `UntagResource` di `aws-external-anthropic` namespace. Tombol tag yang sama dapat mendorong kondisi IAM dan alokasi biaya tanpa konfigurasi tambahan.

Tag-based kontrol akses

Anda dapat mengatur `aws-external-anthropic` tindakan pada nilai tag ruang kerja menggunakan tombol konteks `aws:ResourceTag/<key>` kondisi. Kebijakan berikut mengizinkan inferensi hanya pada ruang kerja yang diberi tag: `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"
        }
      }
    }
  ]
}
```

Kunci penggunaan `aws:RequestTag/<key>` dan `aws:TagKeys` kondisi aktif `TagResource` untuk menerapkan kebijakan penandaan (misalnya, mengharuskan ruang kerja untuk dibawa `CostCenter` dan `Owner` diberi tag). Lihat kebijakan IAM untuk contoh lainnya.

Menggunakan Konsol Claude

Platform Claude di AWS menggunakan Konsol Claude standar di platform.claude.com. Saat Anda masuk dari AWS Console, indikator Akun yang dikelola oleh AWS akan muncul di sidebar Claude Console. Cakupan Konsol ke Platform Claude Anda di organisasi AWS. Ini menyediakan analisis penggunaan, rincian biaya, visibilitas batas tarif, visibilitas ruang kerja, dan halaman untuk mengelola file, Keterampilan Agen, pekerjaan batch, agen, sesi, lingkungan, brankas kredensi, dan penyimpanan memori.

Masuk ke

Akses ke Konsol Claude difederasi melalui AWS IAM. Lihat Mengatur akun Anda untuk alur masuk pertama kali penuh. Singkatnya:

1. Asumsikan peran IAM dengan `aws-external-anthropic:AssumeConsole` izin. Lihat tindakan IAM.
2. Arahkan ke halaman Platform Claude di AWS di AWS Console.
3. Pilih Buka Konsol Claude. AWS Console mengeluarkan JWT dan mengarahkan Anda ke. `platform.claude.com`
4. Saat masuk pertama, Anda akan diminta alamat email; masukkan email kantor Anda. Platform menyediakan pengguna Claude Console Anda tepat waktu.

Dua peran Claude Console tersedia: Admin dan Pengembang. Peran Admin memberikan akses ke semua halaman dan setelan Konsol Claude yang tersedia untuk Platform Claude di AWS. Peran Pengembang memberikan akses baca ke informasi penggunaan, biaya, batas tarif, dan ruang kerja. Hubungi perwakilan akun Anthropic Anda untuk menetapkan peran Admin atau Pengembang ke kepala sekolah.

Halaman yang tersedia

Kolom gateway Via AWS menunjukkan apakah halaman membaca dan menulis data melalui gateway AWS (dan karenanya diatur oleh tindakan IAM). Halaman yang ditandai Tidak membaca metadata tingkat organisasi secara langsung melalui API Antropik dan melewati pemeriksaan tindakan IAM.

Halaman	Available	Melalui gateway AWS	Catatan
Penggunaan	Ya	Tidak	Lihat penggunaan token berdasarkan model, ruang kerja, dan dimensi. Data mungkin memerlukan beberapa menit untuk muncul setelah permintaan.
Biaya	Ya	Tidak	Lihat rincian biaya berdasarkan model dan ruang kerja. AWS Cost Explorer menampilkan item baris CCU agregat.
Batas	Ya	Tidak	Lihat batas tarif (hanya-baca).
Ruang kerja	Ya	Tidak	Lihat ruang kerja per wilayah (hanya-baca).
Berkas	Ya	Ya	Lihat dan kelola file yang diunggah.
Keterampilan	Ya	Ya	Lihat dan kelola Keterampilan Agen.
Batch	Ya	Ya	Lihat dan kelola pekerjaan pemrosesan batch.
Agen	Ya	Ya	Lihat dan kelola definisi agen.
Sesi	Ya	Ya	Lihat sesi agen dan riwayat acara.
Lingkungan	Ya	Ya	Lihat dan kelola konfigurasi kontainer cloud untuk sesi.
Brankas kredensi	Ya	Ya	Lihat dan kelola brankas kredensi untuk otentikasi sesi.
Menyimpan memori	Ya	Ya	Lihat dan kelola memori agen persisten.
Kunci API	Tidak	N/A	Kelola kunci API di AWS Console (Platform Claude di AWS → kunci API). Lihat Otentikasi.
Anggota	Tidak	N/A	Tidak berlaku. AWS IAM mengelola akses.

Halaman	Available	Melalui gateway AWS	Catatan
Penagihan	Tidak	N/A	Tidak berlaku. AWS Marketplace mengelola penagihan dan faktur. Lihat rincian biaya di halaman Biaya.
Kode Claude	Tidak	N/A	Lihat penggunaan Kode Claude di halaman Penggunaan.

Beralih organisasi

Konsol Claude tidak mendukung peralihan organisasi untuk Platform Claude di AWS. Untuk mengakses organisasi lain, keluar dan autentikasi ulang melalui AWS Console menggunakan peran IAM untuk akun AWS organisasi tersebut.

Batas nilai dan kuota

Platform Claude di AWS menetapkan batas tarif Tingkat 1 saat Anda berlangganan. Anthropic mengelola batas tarif secara langsung, bukan melalui sistem kuota AWS.

Batas default

Platform Claude di AWS menggunakan jadwal tingkat standar Anthropic, identik dengan Claude API pihak pertama. Batas tingkat 1 berlaku per ruang kerja. Batas dikumpulkan oleh keluarga model. Misalnya, semua model Opus berbagi satu batas gabungan, semua model Soneta berbagi yang lain, dan semua model Haiku berbagi yang ketiga.

Untuk nilai Tier 1 saat ini (RPM, ITPM, OTPM) dan ambang batas tingkat yang lebih tinggi, lihat Batas nilai di situs web dokumentasi Antropik. Halaman Antropik adalah sumber kebenaran dan diperbarui ketika batas berubah.

Header batas nilai

Setiap respons menyertakan header yang melaporkan status batas tarif Anda saat ini. Header kunci:

- `anthropic-ratelimit-requests-limit`— Permintaan maksimum per menit
- `anthropic-ratelimit-requests-remaining`— Permintaan yang tersisa di jendela saat ini
- `anthropic-ratelimit-requests-reset`— Waktu ketika batas permintaan diatur ulang (RFC 3339)
- `anthropic-ratelimit-tokens-limit`— Token gabungan maksimum (input+output) per menit
- `anthropic-ratelimit-tokens-remaining`— Token gabungan yang tersisa di jendela saat ini
- `anthropic-ratelimit-tokens-reset`— Waktu ketika batas token gabungan diatur ulang (RFC 3339)
- `anthropic-ratelimit-input-tokens-limit`/`-remaining`/`-reset`— Input-token-specific header
- `anthropic-ratelimit-output-tokens-limit`/`-remaining`/`-reset`— Output-token-specific header
- `retry-after`— Pada respons 429, jumlah detik untuk menunggu sebelum mencoba lagi

Lihat Header respons di situs web dokumentasi Antropik untuk set lengkapnya.

Meminta batas yang lebih tinggi

Tidak seperti Claude API pihak pertama, kemajuan tingkat otomatis tidak berlaku pada Platform Claude di AWS. Untuk meminta batas yang lebih tinggi, hubungi perwakilan akun Anthropic Anda dengan ID ruang kerja Anda dan throughput yang diinginkan. Untuk ambang batas tingkat dan detail lainnya, lihat Batas nilai di situs web dokumentasi Antropik.

Kesalahan batas nilai

Bila Anda melebihi batas tarif, API mengembalikan HTTP 429 dengan `rate_limit_error` tipe. Terapkan backoff eksponensial dengan jitter dalam logika coba lagi Anda. `retry-after`Header menunjukkan berapa detik untuk menunggu sebelum mencoba lagi.

Penagihan

Platform Claude di AWS menggunakan AWS Marketplace sebagai backend penagihan untuk pengukuran berbasis konsumsi. Penggunaan pengukur antropik setiap jam melalui AWS Marketplace, dan AWS mengeluarkan faktur bulanan pada tagihan AWS Anda yang ada.

Penagihan hanya tunggakan (Anda membayar untuk apa yang Anda gunakan) dan sebelum pajak (AWS menerapkan pengaturan pajak akun Anda).

Penggunaan dalam mata uang Claude Consumption Units (CCU) seharga $0,01 USD per CCU. Harga CCU tetap dan tidak pernah didiskon. Anthropic menilai penggunaan token Anda dalam USD dengan standar per model, tarif per fitur, menerapkan diskon yang dinegosiasikan, lalu mengonversi hasilnya menjadi CCU seharga $0,01 per CCU. Diskon menghasilkan lebih sedikit CCU yang diukur, bukan harga CCU yang lebih rendah. CCU bukan kredit prabayar; tidak ada saldo atau komitmen CCU. Lihat Harga untuk definisi CCU dan tarif token per model.

Pencatatan dan pemantauan

AWS CloudTrail dapat menangkap semua permintaan ke Platform Claude di AWS. Operasi ruang kerja dan vault dicatat sebagai peristiwa Manajemen secara default. Semua operasi lainnya (inferensi, batch, file, keterampilan, model, profil pengguna, dan Agen Terkelola Claude kecuali brankas) adalah peristiwa Data. Pencatatannya memerlukan konfigurasi eksplisit dan menimbulkan biaya tambahan. CloudTrail Lihat tindakan IAM untuk klasifikasi jenis peristiwa lengkap dan CloudTrail dokumentasi AWS untuk detail konfigurasi.

Saat mengonfigurasi pencatatan peristiwa CloudTrail data, pilih jenis `AWS::AWSExternalAnthropic::Workspace` sumber daya. Ini adalah jenis CloudTrail sumber daya untuk Platform Claude di ruang kerja AWS dan diperlukan untuk menangkap peristiwa bidang data (inferensi, batch, file, dan operasi per ruang kerja lainnya). Cakupan pemilih peristiwa data ke ARN ruang kerja tertentu jika Anda ingin mencatat aktivitas untuk subset ruang kerja, bukan seluruh akun.

Minta ID

Setiap respons menyertakan dua ID permintaan di header respons:

- AWS request ID (**`x-amzn-requestid`**): ID utama, diindeks. CloudTrail Gunakan ini saat menyelidiki permintaan melalui AWS tooling atau saat menghubungi dukungan AWS.
- Anthropic request ID (**`request-id`**): ID sekunder. Gunakan ini saat menghubungi dukungan Antropik.

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

Ruby SDK saat ini tidak mengekspos pengakses respons mentah, jadi ID permintaan tidak dapat dibaca dari header respons di Ruby. Jika Anda memerlukan pencatatan Request-ID, gunakan salah satu bahasa lain di atas, atau tangkap ID di lapisan proxy HTTP.

Anthropic merekomendasikan untuk mencatat aktivitas Anda setidaknya selama 30 hari untuk memahami pola penggunaan dan menyelidiki potensi masalah apa pun.

> **Note**
>
> AWS CloudTrail dikonfigurasi dalam akun AWS Anda. Mengaktifkan logging tidak memberikan AWS atau Anthropic akses ke konten Anda di luar apa yang diperlukan untuk penagihan dan operasi layanan.

Metrik dan dasbor penggunaan

Platform Claude di AWS tidak mempublikasikan metrik penggunaan per ruang kerja ke Amazon. CloudWatch Jumlah token, volume permintaan, dan penggunaan per model tersedia melalui permukaan penggunaan Anthropic:

- Tampilan penggunaan Konsol Claude — saat prinsipal bergabung ke Konsol Claude dengan `aws-external-anthropic:AssumeConsole` (lihat kebijakan IAM), dasbor penggunaan menampilkan volume permintaan, jumlah token, dan kerusakan per model untuk ruang kerja yang dapat diakses. Account-wide tampilan penggunaan memerlukan kemampuan konsol admin.
- Anthropic Usage and Cost API — untuk akses terprogram ke data penggunaan dan biaya, termasuk agregasi per ruang kerja dan per-model, lihat Usage and Cost API dalam dokumentasi Antropik.

Untuk pemantauan operasional (tingkat kesalahan, latensi, header batas tarif), gunakan header respons yang ditampilkan pada setiap permintaan (lihat Batas nilai dan kuota) bersama dengan ID permintaan AWS yang ditangkap. CloudTrail

Alokasi dan pemantauan biaya

Platform Claude pada biaya AWS muncul pada tagihan AWS Anda di bawah Platform Claude pada layanan AWS, dengan dimensi penggunaan per model. Gunakan Laporan Biaya dan Penggunaan AWS (CUR) yang dikombinasikan dengan tag ruang kerja untuk alokasi biaya berbutir halus:

1. Tandai ruang kerja Anda dengan dimensi alokasi yang Anda pedulikan (tim, aplikasi, lingkungan, pusat biaya). Lihat Ruang kerja untuk detail penandaan.
2. Di konsol AWS Billing, aktifkan setiap kunci tag sebagai tag alokasi biaya. Setelah aktivasi, penggunaan terjadi pada atau setelah tanggal aktivasi dibagi dengan tag di CUR.
3. Di CUR atau AWS Cost Explorer, kelompokkan menurut `resourceTags/user:<tag-key>` kolom untuk memecah pembelanjaan berdasarkan tag ruang kerja.

Tag ruang kerja mengalir ke item baris CUR untuk semua Platform Claude pada penggunaan AWS. Tombol tag yang sama mendorong kontrol IAM-based akses dan alokasi biaya. Lihat Tag alokasi biaya dalam dokumentasi AWS Billing untuk langkah-langkah penyiapan dan penundaan aktivasi.

Migrasi dari Amazon Bedrock

Jika saat ini Anda menggunakan Claude di Bedrock, migrasi ke Platform Claude di AWS memerlukan perubahan di seluruh integrasi Anda. Penandatanganan SigV4 tetap didukung, tetapi konteks penandatanganan, URL dasar, format API, ID model, klien dan paket SDK, format streaming, header permintaan, dan ketersediaan wilayah semuanya berubah. Tabel berikut merangkum perbedaannya.

Apa yang berubah

	Amazon Bedrock	Platform Claude di AWS
URL dasar	`bedrock-runtime.{region}.amazonaws.com`	`aws-external-anthropic.{region}.api.aws`
Format API	Batuan Dasar Converse/ InvokeModel	API Pesan Antropik () `/v1/messages`
ID Model	`anthropic.claude-opus-4-7-v1`	`claude-opus-4-7`
Klien SDK	`AnthropicBedrock` /SDK Batuan Dasar	Platform-specific klien (`Anthropic` `AWS` ,,`AnthropicAws` ,`Anthropic` `AwsClient` , dll.), Dalam versi beta
SDK Paket	`anthropic[bedrock]` ,`@anthropic-ai/bedrock-sdk`, dll.	`anthropic[aws]` ,`@anthropic-ai/aws-sdk` , dll. (lihat Menginstal SDK)
Nama layanan SiGv4	`bedrock`	`aws-external-anthropic`
Format streaming	AWS EventStream	SSE (sama seperti Claude API)
Header ruang kerja	Tidak berlaku	`anthropic-workspace-id`diperlukan

	Amazon Bedrock	Platform Claude di AWS
Ketersediaan Wilayah	Beberapa wilayah komersial	Semua wilayah komersial AWS (wilayah keikutsertaan memerlukan keikutsertaan akun)

Apa yang Anda peroleh

- Biasanya akses hari yang sama ke model dan fitur baru, tanpa langkah integrasi mitra yang terpisah
- Keterampilan Agen untuk pembuatan dokumen (PowerPoint, Excel, Word, PDF)
- Eksekusi kode di kotak pasir terkelola Anthropic
- Fitur beta melalui `anthropic-beta` header (lihat Fitur yang saat ini tidak tersedia)
- Claude Console untuk visibilitas kuota dan analisis penggunaan
- Dukungan Antropik Langsung
- Otentikasi kunci API sebagai alternatif untuk SigV4 (lihat Otentikasi)

Apa yang tetap sama

- Autentikasi AWS IAM (SiGv4)
- Penagihan melalui akun AWS
- Pensiun komitmen AWS

Perangkap migrasi

> **Warning**
>
> Aktifkan federasi identitas web keluar terlebih dahulu. Jika akun AWS Anda sebelumnya tidak menggunakan Platform Claude di AWS, Anda harus mengaktifkan federasi identitas web keluar satu kali per akun sebelum mengajukan permintaan. Tanpa langkah ini, semua permintaan gagal dengan kesalahan federasi. Langkah ini tidak diperlukan untuk Bedrock.

> **Warning**
>
> Zero Data Retention (ZDR) adalah opt-in di Platform Claude di AWS. Di Bedrock, AWS adalah pemroses data dan Anthropic tidak menyimpan input atau output inferensi; Program ZDR Anthropic tidak berlaku di sana. Di Platform Claude di AWS, Anthropic adalah pemroses data, dan ZDR mengikuti model Claude API pihak pertama: tersedia berdasarkan permintaan melalui perwakilan akun Anthropic Anda. Konfirmasikan pendaftaran ZDR sebelum memigrasikan beban kerja produksi yang bergantung pada jaminan penyimpanan data.

Pertimbangan komersial

- Ketentuan layanan: Ketentuan Layanan AWS mengatur Platform Claude di AWS. Ketentuan Layanan Komersil Anthropic, Adendum Pemrosesan Data, dan Kebijakan Penggunaan juga berlaku.
- Diskon dan penawaran pribadi: Diskon yang dinegosiasikan dan penawaran pribadi AWS Marketplace tidak ditransfer secara otomatis antara Platform Bedrock dan Claude di AWS. Bekerja sama dengan perwakilan akun Anthropic Anda untuk menyiapkan persyaratan komersial untuk Platform Claude di AWS.

Kebijakan IAM

Platform Claude di AWS terintegrasi dengan AWS IAM untuk kontrol akses. Anda memberikan atau menolak akses ke tindakan API tertentu pada ruang kerja tertentu menggunakan sintaks kebijakan IAM standar.

Nama layanan SiGv4 dan namespace tindakan IAM adalah. `aws-external-anthropic` Tindakan mengikuti pola `aws-external-anthropic:<Action>` (misalnya,`aws-external-anthropic:CreateInference`).

Contoh: tolak inferensi batch

Kebijakan berikut memungkinkan inferensi real-time sambil memblokir pemrosesan batch, persyaratan umum untuk beban ZDR-sensitive kerja:

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

```
}
```

`GetBatchInference`Tindakan ini mengotorisasi rute metadata batch dan rute hasil batch. Menyangkalnya, di samping itu`ListBatchInferences`, memblokir pembacaan dan pencacahan batch.

`Allow`Pernyataan tersebut menyebutkan spesifik `Get*` dan `List*` tindakan daripada menggunakan wildcard. Wildcard akan memberikan `GetFile` (yang mengunduh byte file) dan pembacaan lain yang mungkin tidak Anda maksudkan; `Deny` mengganti `Allow` apa pun, tetapi formulir eksplisit adalah pola yang lebih aman untuk dimodelkan.

Contoh: inferensi sinkron pada satu ruang kerja

Memberikan izin minimal untuk prinsipal IAM yang menjalankan inferensi terhadap satu ruang kerja produksi:

```
{
  "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**
>
> `List*`Wildcard dalam kebijakan ini juga cocok`ListWorkspaces`, yang memiliki cakupan akun. Batasan ARN ruang kerja menyaringnya secara diam-diam, sehingga kebijakan ini tidak mengizinkan daftar ruang kerja. Jika akun layanan Anda perlu menghitung ruang kerja, tambahkan pernyataan terpisah `Allow` untuk with. `ListWorkspaces` `Resource: "*"`

Kebijakan ini mengasumsikan autentikasi AWS SiGv4. Jika prinsipal mengautentikasi dengan kunci API, berikan juga `aws-external-anthropic:CallWithBearerToken` (lihat Otentikasi).

Contoh: isolasi ruang kerja per pelanggan

Membatasi peran ke satu ruang kerja:

```
{
  "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

`aws-external-anthropic:*`Wildcard dalam pernyataan pertama menyertakan tindakan cakupan akun (`CreateWorkspace`,) yang disaring oleh batasan `ListWorkspaces` ARN ruang kerja secara diam-diam. Ini konsisten dengan maksud “isolasi” — peran tidak dapat membuat atau menghitung ruang kerja — tetapi kebijakan tersebut berisi izin yang tidak berpengaruh. Lihat Otomatisasi penyediaan untuk pola cakupan akun.
Pernyataan kedua memberikan `CallWithBearerToken` dan `AssumeConsole` pada semua sumber daya karena keduanya adalah tindakan tanpa rute yang tidak mengikat ARN ruang

kerja. Hilangkan pernyataan kedua jika peran hanya menggunakan SiGv4 dan tidak pernah berfederasi ke Konsol Claude.

Contoh: penguncian fitur untuk ruang kerja ZDR-sensitive

Memblokir pemrosesan batch dan unggahan file di ruang kerja tertentu sambil membiarkan inferensi sinkron tersedia. Berguna saat ruang kerja menangani data Zero Data Retention (ZDR) yang tidak boleh bertahan di sisi server. Lampirkan kebijakan ini di samping kebijakan Izinkan seperti `AnthropicLimitedAccess` atau contoh ruang kerja tunggal di atas; dengan sendirinya, Deny-only kebijakan tidak memberikan izin:

```
{
  "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

Ini menyangkal hanya memblokir pembuatan. Tindakan file dan batch lainnya tidak ditolak kecuali Anda mencantumkannya juga. Untuk penguncian lengkap di mana ruang kerja tidak boleh menyimpan file atau batch, tolak juga,,`aws-external-anthropic:GetFile`,,`aws-external-anthropic:ListFiles`, `aws-external-anthropic:DeleteFile` `aws-external-anthropic:GetBatchInferenceaws-external-anthropic:ListBatchInferences`, `aws-external-anthropic:CancelBatchInference` dan. `aws-external-anthropic:DeleteBatchInference`

Contoh: otomatisasi penyediaan

Memberikan CI/CD peran tindakan yang diperlukan untuk membuat dan mengelola ruang kerja, tanpa izin inferensi apa pun:

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

`CreateWorkspace`dan `ListWorkspaces` merupakan operasi cakupan akun. Menentukan ARN ruang kerja pada tindakan ini tidak berpengaruh; gunakan. `Resource: "*"`

Kebijakan terkelola

AWS menyediakan kebijakan terkelola untuk pola akses umum:

- **`AnthropicFullAccess`**: Hibah `aws-external-anthropic:*` pada semua sumber daya.
- **`AnthropicReadOnlyAccess`**: Hibah`Get*`,`List*`, dan `CallWithBearerToken` pada semua sumber daya.
- **`AnthropicInferenceAccess`**: Memberikan ReadOnly tindakan ditambah tindakan inferensi (`CreateInference`,,, `CreateBatchInference` `CancelBatchInferenceDeleteBatchInference`,`CountTokens`) pada semua sumber daya.
- **`AnthropicLimitedAccess`**: Memberikan `AnthropicInferenceAccess` tindakan ditambah semua tindakan Agen Terkelola Claude (agen, sesi, lingkungan, brankas, penyimpanan memori) di semua sumber daya.

- **`AnthropicSelfHostedEnvironmentAccess`**: Memberikan izin yang dibutuhkan pekerja kotak pasir yang dihosting sendiri untuk melakukan polling dan memproses item kerja lingkungan, membaca lingkungan terkait, sesi, dan sumber daya keterampilan, dan memperbarui status sesi. Lampirkan kebijakan ini ke peran IAM yang diasumsikan oleh pekerja lingkungan yang dihosting sendiri. `AnthropicSelfHostedEnvironmentAccess`adalah default peran tunggal yang direkomendasikan; jika Anda menjalankan poller kerja dan kotak pasir per sesi di bawah prinsip IAM terpisah, Anda dapat membagi izin ini di dua kebijakan kustom yang lebih sempit untuk hak istimewa yang paling sedikit.

`AnthropicInferenceAccess`adalah kebijakan terkelola tersempit yang cukup untuk menjalankan inferensi. Melalui wildcard `Get*` dan `List*` wildcard, ia memberikan akses baca ke setiap sumber daya API di namespace, termasuk unduhan konten file `GetFile` dan konten memori. `GetMemoryStore` Itu tidak memberikan pembuatan atau penghapusan file atau keterampilan, manajemen profil pengguna, mutasi ruang kerja, atau federasi konsol.

> **Note**
>
> `AnthropicReadOnlyAccess`,`AnthropicInferenceAccess`, dan `AnthropicLimitedAccess` jangan mengabulkan`AssumeConsole`. Kepala sekolah yang perlu berfederasi ke Konsol Claude memerlukan hibah terpisah untuk `aws-external-anthropic:AssumeConsole` — baik melalui atau kebijakan khusus. `AnthropicFullAccess` Lihat Federasi ke Konsol Claude.

> **Note**
>
> `CreateInference`dan `CreateBatchInference` merupakan tindakan yang terpisah. Menyangkal yang satu tidak menghalangi yang lain. Jika Anda bermaksud mencegah semua panggilan model, tolak keduanya.

Federasi ke Konsol Claude

`aws-external-anthropic:AssumeConsole`memungkinkan federasi utama IAM ke Konsol Anthropic-operated Claude. Akses di dalam konsol masih diatur oleh IAM untuk sebagian besar operasi, tetapi sebagian dari operasi admin - terutama tampilan penggunaan yang tidak memiliki API IAM yang sesuai - diberi gerbang pada kemampuan yang digabungkan dengan prinsipal.

Ada dua kemampuan:

- **`developer`**— mengizinkan pengoperasian Platform Claude pada kebutuhan pengembang AWS untuk pekerjaan sehari-hari: menjalankan inferensi dari konsol, membaca data ruang kerja, melihat penggunaan pribadi.
- **`admin`**— Selain itu mengizinkan operasi konsol khusus admin, termasuk tampilan penggunaan seluruh akun dan pengaturan administratif yang tidak muncul melalui tindakan IAM.

Kontrol kemampuan mana yang dapat diminta oleh prinsipal dengan tombol `aws-external-anthropic:Capability` kondisi pada `AssumeConsole` tindakan:

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

`AnthropicFullAccess`hibah `AssumeConsole` tanpa batasan kemampuan. Untuk hibah yang lebih sempit — akses konsol khusus pengembang, atau akses khusus admin — lampirkan kebijakan khusus yang mencakup kondisi seperti yang ditunjukkan di atas. `AssumeConsole` `aws-external-anthropic:Capability`

Memanggil dengan token pembawa

`aws-external-anthropic:CallWithBearerToken`mengotorisasi jalur SigV4-free permintaan yang digunakan saat kunci API disajikan sebagai token pembawa. Setiap prinsipal yang mengautentikasi dengan kunci API memerlukan tindakan ini di ruang kerja target. Ini berlaku apakah Anda menggunakan `ANTHROPIC_AWS_API_KEY` atau mengatur `Authorization: Bearer` header secara langsung.

Ini diperlukan untuk pemanggil kunci API selain tindakan inferensi (`CreateInference`,`CreateBatchInference`, dll.). Tanpa`CallWithBearerToken`, permintaan kunci API ditolak sebelum mencapai pemeriksaan otorisasi inferensi. Penelepon SiGv4 tidak memerlukan tindakan ini.

`AnthropicReadOnlyAccess`,`AnthropicInferenceAccess`,`AnthropicLimitedAccess`, dan `AnthropicFullAccess` semua termasuk`CallWithBearerToken`. Jika Anda menulis kebijakan khusus untuk akses kunci API, tambahkan secara eksplisit.

Federasi identitas web keluar (diperlukan untuk akses konsol)

Konsol Claude berjalan di infrastruktur Antropik, bukan AWS. Ketika panggilan utama IAM`AssumeConsole`, AWS STS mengeluarkan token identitas web yang dicakup oleh audiens Anthropic; konsol Anthropic kemudian menerima token itu dan menetapkan sesi federasi. Agar ini berfungsi, akun AWS harus mengizinkan federasi identitas web keluar kepada `aws-external-anthropic` audiens.

Prinsipal yang memanggil `AssumeConsole` memerlukan izin STS berikut selain tindakan: `aws-external-anthropic:AssumeConsole`

- **`sts:GetWebIdentityToken`**— memungkinkan penerbitan token identitas web yang dikonsumsi konsol Anthropic.
- **`sts:TagGetWebIdentityToken`**— memungkinkan melampirkan tag sesi ke token identitas web. Platform Claude di AWS menggunakan tag ini untuk menyampaikan kapabilitas kepala sekolah dan konteks ruang kerja ke konsol.

Sertakan keduanya dalam kebijakan kepercayaan dari peran apa pun yang diasumsikan oleh pengguna konsol, atau dalam inline/managed kebijakan yang dilampirkan pada identitas pengguna yang menelepon `AssumeConsole` secara langsung. `AnthropicFullAccess`mencakup kedua tindakan STS. Lingkungan yang menolak `sts:*` di SCP atau tingkat batas izin harus secara eksplisit mengizinkan kedua tindakan ini agar federasi konsol berhasil.

Tindakan IAM untuk Platform Claude di AWS

Referensi tindakan IAM untuk mengontrol akses ke Platform Claude di AWS melalui kebijakan AWS.

Platform Claude di AWS menggunakan AWS IAM untuk kontrol akses. Setiap rute API memetakan ke tindakan IAM di `aws-external-anthropic` namespace. Halaman ini mencantumkan semua tindakan, rute yang diotorisasi oleh setiap tindakan, dan kebijakan terkelola yang tersedia untuk pola akses umum. Untuk penyiapan dan autentikasi platform, lihat Apa itu Platform Claude di AWS? .

Detail layanan

Awalan layanan IAM	`aws-external-anthropic`
Jenis sumber daya	`workspace`
Ruang kerja ARN	`arn:aws:aws-external-anthropic:{region}:{account-id}:workspace/{workspace-id}`

Wilayah ARN selalu dihuni dan cocok dengan wilayah tempat ruang kerja terikat. Segmen sumber daya adalah ID ruang kerja yang diberi tag (`wrkspc_…`), nilai yang sama yang Anda berikan di `anthropic-workspace-id` header.

Tindakan

Layanan ini mendefinisikan 65 tindakan di 15 grup. Tindakan mengikuti `VerbNoun` konvensi AWS dan menggunakan disiplin kata kerja sehingga `Get*` `List*` wildcard menghasilkan batas hanya-baca yang bersih.

Inferensi

Tindakan	Rute diotorisasi
`CreateInference`	`POST /v1/messages`
`CountTokens`	`POST /v1/messages/count_tokens`

Pemrosesan batch

Tindakan	Rute diotorisasi
`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}`

Model

Tindakan	Rute diotorisasi
`GetModel`	`GET /v1/models/{id}`
`ListModels`	`GET /v1/models`

Berkas

Tindakan	Rute diotorisasi
`CreateFile`	`POST /v1/files`
`GetFile`	`GET /v1/files/{id}  GET /v1/files/{id}/content`
`ListFiles`	`GET /v1/files`
`DeleteFile`	`DELETE /v1/files/{id}`

> **Note**
>
> `GetFile`mengotorisasi unduhan metadata dan konten. Prinsipal dengan akses hanya-baca dapat mengunduh byte file, bukan hanya daftar file.

Keterampilan

Tindakan	Rute diotorisasi
`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**
>
> Menghapus peta versi keterampilan individu ke`UpdateSkill`, bukan`DeleteSkill`. Kebijakan yang menyangkal `aws-external-anthropic:Delete*` masih mengizinkan penghapusan versi. Tolak `UpdateSkill` dan `CreateSkill` juga jika Anda perlu mencegah mutasi keterampilan apa pun.

Profil pengguna

Tindakan	Rute diotorisasi
`CreateUserProfile`	`POST /v1/user_profiles`
`GetUserProfile`	`GET /v1/user_profiles/{id}`
`ListUserProfiles`	`GET /v1/user_profiles`
`UpdateUserProfile`	`POST /v1/user_profiles/{id}`

> **Warning**
>
> Pencocokan aksi IAM tidak peka huruf besar/kecil. Wildcard `aws-external-anthropic:*File` cocok`CreateFile`,`GetFile`, dan`DeleteFile`, tetapi tidak cocok `ListFiles` (yang berakhir dengan “file”, bukan “file”). Ini juga terlalu cocok`CreateUserProfile`,`GetUserProfile`, dan `UpdateUserProfile` karena “Profil” berakhir dengan “file”. Jika Anda bermaksud memberikan atau menolak hanya tindakan API File, sebutkan secara eksplisit (`CreateFile`,, `GetFile``ListFiles`,`DeleteFile`) daripada menggunakan pola sufiks. `*File`

Agen

Definisi agen untuk Agen Terkelola Claude.

Tindakan	Rute diotorisasi
`CreateAgent`	Agen membuat rute
`GetAgent`	Agen membaca rute
`ListAgents`	Rute daftar agen
`UpdateAgent`	Rute pembaruan agen
`ArchiveAgent`	Rute arsip agen

Sesi

Sesi agen dan riwayat acara.

Tindakan	Rute diotorisasi
`CreateSession`	Sesi membuat rute
`GetSession`	Rute membaca sesi
`ListSessions`	Rute daftar sesi
`UpdateSession`	Rute pembaruan sesi
`ArchiveSession`	Rute arsip sesi
`DeleteSession`	Sesi menghapus rute

Lingkungan

Konfigurasi kontainer cloud untuk sesi.

Tindakan	Rute diotorisasi
`CreateEnvironment`	Lingkungan membuat rute
`GetEnvironment`	Lingkungan membaca rute
`ListEnvironments`	Rute daftar lingkungan
`UpdateEnvironment`	Rute pembaruan lingkungan
`ArchiveEnvironment`	Rute arsip lingkungan
`DeleteEnvironment`	Lingkungan menghapus rute
`ProcessEnvironmentWork`	Self-hosted rute pekerja lingkungan

Vault

Credential vaults untuk otentikasi sesi.

Tindakan	Rute diotorisasi
`CreateVault`	Vault membuat rute
`GetVault`	Vault membaca rute
`ListVaults`	Rute daftar Vault
`UpdateVault`	Rute pembaruan Vault
`ArchiveVault`	Rute arsip vault
`DeleteVault`	Vault menghapus rute

> **Note**
>
> Operasi Vault diklasifikasikan sebagai peristiwa CloudTrail Manajemen (bukan peristiwa Data) karena vault menyimpan kredensil dan mendapat manfaat dari pencatatan audit default. Operasi Agen Terkelola Claude lainnya (agen, sesi, lingkungan, penyimpanan memori) adalah peristiwa Data.

Menyimpan memori

Memori agen persisten.

Tindakan	Rute diotorisasi
`CreateMemoryStore`	Toko memori membuat rute
`GetMemoryStore`	Rute baca toko memori
`ListMemoryStores`	Rute daftar toko memori
`UpdateMemoryStore`	Rute pembaruan toko memori

Tindakan	Rute diotorisasi
`ArchiveMemoryStore`	Rute arsip penyimpanan memori
`DeleteMemoryStore`	Rute hapus toko memori

> **Note**
>
> `GetMemoryStore`membaca isi memori. Oleh karena itu, `Get*` wildcard dalam kebijakan terkelola atau kustom memberikan akses baca memori. Kebijakan lingkup secara eksplisit jika konten memori harus dibatasi.

Webhook

Pemberitahuan peristiwa siklus hidup untuk sesi.

Tindakan	Rute diotorisasi
`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`

Workspace

Tindakan	Rute diotorisasi
`CreateWorkspace`	`POST /v1/organizations/workspaces`

Tindakan	Rute diotorisasi
`GetWorkspace`	`GET /v1/organizations/workspaces/{id}`
`ListWorkspaces`	`GET /v1/organizations/workspaces`
`UpdateWorkspace`	`POST /v1/organizations/workspaces/{id}`
`ArchiveWorkspace`	`POST /v1/organizations/workspaces/{id}/archive`

Ruang kerja default disediakan saat mendaftar; lihat Ruang kerja.

Autentikasi

Tindakan	Rute diotorisasi
`CallWithBearerToken`	(tidak ada)

`CallWithBearerToken`adalah izin auth-layer yang mengotorisasi prinsipal untuk mengautentikasi melalui kunci API (token pembawa) daripada AWS SIgV4. Itu tidak memetakan ke rute. Berikan di samping tindakan yang dipetakan rute yang Anda ingin dilakukan oleh pemegang kunci API.

Akses konsol

Tindakan	Rute diotorisasi
`AssumeConsole`	(tidak ada)

`AssumeConsole`memberi wewenang kepada prinsipal untuk membuka Konsol Claude untuk Platform Claude di ruang kerja AWS melalui alur federasi AWS Console. Itu tidak memetakan ke rute. Berikan kepada kepala sekolah yang seharusnya dapat mengklik Buka Konsol Claude di Platform Claude pada halaman layanan AWS di AWS Console. Kunci `aws-external-anthropic:Capability` kondisi pada `AssumeConsole` tindakan mengontrol kemampuan konsol (pengembang atau admin) mana yang dapat diminta oleh prinsipal. Lihat kebijakan IAM untuk detailnya dan Menggunakan Konsol Claude untuk alur masuk.

> **Warning**
>
> `aws-external-anthropic:AssumeConsole`mengeluarkan sesi Konsol Claude yang berlangsung hingga 12 jam, terlepas dari durasi sesi pemanggil sendiri. Penelepon yang sesi IAMNYA lebih pendek dari 12 jam masih bisa mendapatkan sesi konsol yang melebihi kredensi sumbernya. Batasi izin ini untuk kepala sekolah yang memerlukan akses Konsol Claude, dan segera cabut jika tidak diperlukan lagi.

> **Note**
>
> Tindakan yang dilakukan di dalam Konsol Claude setelah federasi tidak direkam di AWS CloudTrail atau dapat diatribusikan melalui IAM. Ini termasuk aktivitas global dengan cakupan ruang kerja seperti melihat laporan penggunaan. Jika Anda memerlukan jejak audit untuk aktivitas konsol, gunakan log audit yang tersedia di Konsol Claude, bukan. CloudTrail

Route-to-action pemetaan

Tabel berikut mencantumkan setiap rute di Platform Claude di AWS dan tindakan IAM yang diperlukan untuk memanggilnya. Tindakan IAM rute stabil juga mengotorisasi permintaan yang menggunakan header. `anthropic-beta` CloudTrail mengklasifikasikan setiap tindakan sebagai peristiwa Data (volume tinggi, operasi data-plane) atau peristiwa Manajemen (operasi bidang kontrol).

Metode	Rute	Tindakan IAM	CloudTrail jenis acara
`POST`	`/v1/messages`	`CreateInference`	Data
`POST`	`/v1/messages/count_tokens`	`CountTokens`	Data
`POST`	`/v1/messages/batches`	`CreateBatchInference`	Data
`GET`	`/v1/messages/batches`	`ListBatchInferences`	Data
`GET`	`/v1/messages/batches/{id}`	`GetBatchInference`	Data

Metode	Rute	Tindakan IAM	CloudTrail jenis acara
GET	/v1/messages/batches/{id}/results	GetBatchInference	Data
POST	/v1/messages/batches/{id}/cancel	CancelBatchInference	Data
DELETE	/v1/messages/batches/{id}	DeleteBatchInference	Data
GET	/v1/models	ListModels	Data
GET	/v1/models/{id}	GetModel	Data
POST	/v1/files	CreateFile	Data
GET	/v1/files	ListFiles	Data
GET	/v1/files/{id}	GetFile	Data
GET	/v1/files/{id}/content	GetFile	Data
DELETE	/v1/files/{id}	DeleteFile	Data
POST	/v1/skills	CreateSkill	Data
GET	/v1/skills	ListSkills	Data
GET	/v1/skills/{id}	GetSkill	Data
DELETE	/v1/skills/{id}	DeleteSkill	Data
POST	/v1/skills/{id}/versions	UpdateSkill	Data
GET	/v1/skills/{id}/versions	GetSkill	Data
GET	/v1/skills/{id}/versions/{version}	GetSkill	Data

Metode	Rute	Tindakan IAM	CloudTrail jenis acara
GET	/v1/skills/{id}/versions/{version}/content	GetSkill	Data
DELETE	/v1/skills/{id}/versions/{version}	UpdateSkill	Data
POST	/v1/user_profiles	CreateUserProfile	Data
GET	/v1/user_profiles	ListUserProfiles	Data
GET	/v1/user_profiles/{id}	GetUserProfile	Data
POST	/v1/user_profiles/{id}	UpdateUserProfile	Data
POST	/v1/organizations/workspaces	CreateWorkspace	Manajemen
GET	/v1/organizations/workspaces	ListWorkspaces	Manajemen
GET	/v1/organizations/workspaces/{id}	GetWorkspace	Manajemen
POST	/v1/organizations/workspaces/{id}	UpdateWorkspace	Manajemen
POST	/v1/organizations/workspaces/{id}/archive	ArchiveWorkspace	Manajemen

Rute Agen Terkelola Claude (agen, sesi, lingkungan, brankas, penyimpanan memori) mengikuti pola route-to-action yang sama; untuk pemetaan per rute lengkap, lihat referensi tindakan IAM Anthropic.
Rute Vault adalah peristiwa Manajemen; agen, sesi, lingkungan, dan rute penyimpanan memori adalah peristiwa Data.

Rute yang tidak ada di Platform Claude di AWS ditolak di gateway secara default.

Lihat juga

- Kebijakan IAM misalnya kebijakan dan kebijakan terkelola
- Panduan Pengguna AWS IAM untuk sintaks kebijakan IAM dan logika evaluasi
- Panduan CloudTrail Pengguna AWS untuk konfigurasi pencatatan audit

Riwayat dokumen

Tabel berikut menjelaskan perubahan penting pada Platform Claude di Panduan Pengguna AWS.

Ubah	Deskripsi	Date
Publikasi awal	Menerbitkan Platform Claude di Panduan Pengguna AWS.	7 Mei 2026

Terjemahan disediakan oleh mesin penerjemah. Jika konten terjemahan yang diberikan bertentangan dengan versi bahasa Inggris aslinya, utamakan versi bahasa Inggris.