Guide de l’utilisateur

# Claude Platform sur AWS

# Claude Platform sur AWS: Guide de l’utilisateur

# Table of Contents

# Qu'est-ce que la plateforme Claude sur AWS ?

Accédez aux fonctionnalités complètes de la plateforme Claude via AWS avec Anthropic-managed infrastructure.

Claude Platform sur AWS vous offre l'expérience complète de la plateforme Anthropic, y compris l'API Messages, les compétences des agents, l'exécution du code et les fonctionnalités bêta, accessibles via votre compte AWS. Contrairement à Amazon Bedrock, où AWS gère la pile d'inférence, Anthropic exploite Claude Platform sur AWS. AWS fournit la couche d'authentification (SigV4 ou clé API), le contrôle IAM-based d'accès et l'intégration de la facturation via AWS Marketplace.

> **Note**
>
> Les kits de développement logiciel Anthropic prennent en charge la plateforme Claude sur AWS. Pour connaître la disponibilité des clients par langue, consultez la documentation des SDK d'Anthropic Client.

## Comment fonctionne l'intégration de la plateforme

Les modèles Claude fonctionnent sur Anthropic-managed l'infrastructure. Il s'agit d'une intégration commerciale pour la facturation et l'accès via AWS. Anthropic et AWS agissent tous deux en tant que processeurs de données indépendants. Les conditions de service AWS régissent la plateforme Claude sur AWS. Les conditions d'utilisation commerciales, l'addendum relatif au traitement des données et la politique d'utilisation d'Anthropic s'appliquent également.

Notez les caractéristiques opérationnelles suivantes. Les données ne peuvent pas résider dans AWS. L'inférence peut être acheminée vers le cloud principal d'Anthropic. Les sous-services peuvent être modifiés sans préavis. Définissez le `inference_geo` paramètre par demande pour attribuer l'inférence à une zone géographique spécifique. Voir Résidence des données pour plus de détails.

La plateforme Claude sur AWS suit la même politique de conservation des données que l'API Claude propriétaire. Zero Data Retention (ZDR) est disponible sur demande. Contactez le représentant de votre compte Anthropic pour l'activer pour votre compte.

# Caractéristiques de la plateforme Claude sur AWS

La plateforme Claude sur AWS fournit les fonctionnalités suivantes :

- Same-day accès aux modèles et aux fonctionnalités — Les nouveaux modèles Claude et fonctionnalités de l'API sont disponibles le jour même de leur lancement sur l'API Claude propriétaire.
- Compétences des agents — Utilisez les compétences prédéfinies pour la génération de documents (ExcelPowerPoint, Word, PDF) et créez des compétences personnalisées.
- Exécution du code — Exécutez le code dans le sandbox géré d'Anthropic.
- Pensée étendue — Activez la réflexion étendue pour les tâches de raisonnement complexes.
- Streaming et traitement par lots : utilisez le streaming SSE en temps réel ou soumettez des demandes par lots pour des charges de travail à haut débit.
- Mise en cache rapide : outils de mise en cache, instructions du système et historique des messages pour réduire la latence et les coûts.
- API de fichiers — Téléchargez et référencez des fichiers dans toutes les demandes.
- Intégration à AWS IAM : contrôlez l'accès grâce aux politiques IAM standard et à l'authentification SigV4.
- Facturation AWS unifiée : payez par le biais de votre compte AWS existant grâce au comptage basé sur la consommation (Marketplace en tant que backend de facturation).

Pour obtenir la liste complète des fonctionnalités prises en charge, consultez la section Support des fonctionnalités.

# Comment est organisé ce guide

Ce guide couvre les sujets suivants :

- Mise en route : configuration du compte, prérequis, authentification et installation du SDK.
- Utilisation de l'API : modèles disponibles, envoi de demandes et prise en charge des fonctionnalités.
- Gestion de votre environnement : résidence des données, espaces de travail, console Claude, limites de débit et facturation.
- Surveillance et opérations : CloudTrail journalisation et suivi des identifiants de demande.

- Migration — Migration d'Amazon Bedrock vers Claude Platform sur AWS.
- Sécurité et contrôle d'accès : référence des politiques et actions IAM.

# Claude Platform sur AWS contre Amazon Bedrock

Les deux offres vous permettent d'utiliser Claude via AWS, mais elles diffèrent considérablement en termes d'architecture, de surface d'API et de disponibilité des fonctionnalités.

| | Claude Platform sur AWS | Amazon Bedrock |
|---|---|---|
| Qui gère la pile | Anthropic | AWS |
| Surface de l'API | API de messages anthropiques () `/v1/messages` | API Bedrock (Converse/) InvokeModel |
| Disponibilité des fonctionnalités | Same-day dans le rôle de : Claude API | Dépend du calendrier d'intégration d'AWS |
| Compétences des agents | Available | Non disponible (nécessite l'exécution de code) |
| Fonctionnalités bêta | Transférer avec des `anthropic-beta` en-têtes (voir Support des fonctionnalités) | Nécessite le support AWS |
| Authentification | AWS IAM/SigV4 ou clé d'API | AWS IAM/SigV4 ou jeton porteur (SDK C#, Go et Java uniquement) |
| URL de base | `aws-external-anthropic.{region}.api.aws` | `bedrock-runtime.{region}.amazonaws.com` |
| Client du SDK | Platform-specific classe client (par exemple, `AnthropicAWS` en Python), en version bêta | `AnthropicBedrock` /SDK Bedrock |
| Console | Claude Console (`platform.claude.com` accès via la console AWS) | Console Bedrock |
| Limites de taux et quotas | Géré par Anthropic | Géré par AWS |

| | Claude Platform sur AWS | Amazon Bedrock |
|---|---|---|
| Processeurs de données | Anthropic et AWS | AWS |

Si vous avez besoin de AWS-operated Claude avec l'API Bedrock, consultez Amazon Bedrock.

> **Important**
>
> Anthropic fournit la plateforme Claude sur AWS en tant qu'offre tierce. Il n'est pas couvert par les programmes de conformité, les certifications ou les rapports d'audit standard d'AWS (tels que l'éligibilité aux normes SOC, ISO ou HIPAA). Les clients sont seuls responsables de faire preuve de diligence raisonnable pour s'assurer que cette offre tierce répond à leurs exigences réglementaires, légales et de conformité.

Quand choisir Bedrock : choisissez Amazon Bedrock si votre organisation a besoin d' AWS-operated inférence, d'AWS comme seul processeur de données ou d'une couverture dans le cadre des programmes de conformité d'AWS (y compris l'éligibilité à FedRAMP High, IL4, IL5, SOC, ISO et HIPAA). Bedrock fonctionne entièrement sur une AWS-controlled infrastructure avec AWS comme opérateur.

# configurer votre compte ;

La configuration de Claude Platform sur AWS comporte quatre phases : vous inscrire dans la console AWS, configurer votre organisation Anthropic, noter l'identifiant de votre espace de travail et vous connecter à la console Claude.

> **Note**
>
> L'inscription via la console AWS permet de créer une nouvelle organisation Anthropic liée à votre compte AWS. Cette organisation est distincte de toutes les organisations que votre entreprise possède déjà avec Anthropic, y compris les organisations Claude Enterprise achetées via AWS Marketplace. Les clés d'API, les espaces de travail et les paramètres de la console Claude d'une organisation Anthropic propriétaire ne sont pas reportés.
> Si vous avez déjà une offre privée Amazon Bedrock, contactez votre responsable de compte Anthropic ou AWS avant de vous inscrire afin que votre discount s'applique dès votre première demande. Les remises ne peuvent pas être appliquées rétroactivement à l'utilisation effectuée avant l'acceptation de votre offre privée. Voir les offres privées.

## Étape 1 : Inscrivez-vous dans la console AWS

1. Ouvrez la console AWS et accédez à la page de service Claude Platform on AWS.
2. Choisissez S'inscrire.
3. Sélectionnez Continuer.

La page affiche une bannière Sign-up en cours de réalisation. Restez sur la page. Sign-up prend quelques minutes pendant qu'AWS gère l'abonnement AWS Marketplace pour vous, puis vous redirige automatiquement.

Si votre organisation dispose d'une offre privée d'Anthropic, la console la recherche et vous invite à l'accepter sur AWS Marketplace. Voir les offres privées pour plus de détails.

> **Note**
>
> Si vous utilisez la plateforme Claude sur AWS, votre contenu (tel que les invites et les réponses) est traité par Anthropic en dehors d'AWS. Consultez les politiques d'utilisation des

données d'Anthropic pour plus de détails sur la manière dont le contenu et les métadonnées sont traités et stockés.

## Étape 2 : Configurez votre organisation Anthropic

Une fois l'inscription terminée, vous êtes redirigé vers`platform.claude.com/partner-signup`.

1. Entrez l'adresse e-mail du propriétaire de votre organisation et choisissez Commencer.
2. Consultez cette boîte de réception pour trouver un lien de configuration et suivez-le. Si votre navigateur affiche une page de connexion en tant que compte différent, choisissez Déconnexion et continuez.
3. Remplissez le formulaire de détails de l'organisation (nom de l'organisation, type d'entité, pays, utilisation prévue) et choisissez Compléter la configuration.

L'achèvement de la configuration crée votre organisation Anthropic. Les conditions de service AWS régissent la plateforme Claude sur AWS. Les conditions d'utilisation commerciales, l'addendum relatif au traitement des données et la politique d'utilisation d'Anthropic s'appliquent également. La page de service de la console AWS affiche désormais une navigation vers la gauche avec l'accueil, les clés d'API, Quickstart et les espaces de travail.

## Étape 3 : notez l'identifiant de votre espace de travail

Un espace de travail par défaut est mis en service automatiquement lorsque vous vous inscrivez. Des espaces de travail supplémentaires peuvent être créés par région ; consultez la section Espaces de travail pour plus de détails sur la liaison entre les régions, la gestion de l'espace de travail et l'étendue des ressources IAM.

Trouvez l'ID de l'espace de travail sous Espaces de travail sur la page de service AWS Console Claude Platform on AWS ou dans la console Claude (voir Utilisation de la console Claude). Les identifiants d'espace de travail utilisent le format `wrkspc_` suivi d'un identifiant alphanumérique.

## Étape 4 : Connectez-vous à la console Claude

L'accès à la console Claude est fédéré via AWS IAM :

1. Assumez un rôle IAM avec l'`aws-external-anthropic:AssumeConsole`autorisation. Consultez la section Actions IAM.
2. Sur la page du service Claude Platform on AWS, sélectionnez Sign in. La console AWS émet un JWT et vous redirige vers. `platform.claude.com`
3. Lors de votre première connexion, vous êtes invité à saisir une adresse e-mail. Entrez votre adresse e-mail professionnelle. La plateforme approvisionne votre utilisateur de la console Claude juste à temps.

Lorsque vous êtes connecté via la console AWS, la console Claude s'étend à votre organisation Claude Platform on AWS. Un indicateur de compte géré par AWS apparaît dans la barre latérale de la console Claude.

## Résolution des problèmes de configuration du compte

- « Sign-up  échec : échec de l'activation OutboundWebIdentityFederation » : si vous voyez cette bannière rouge lors de la première soumission, cliquez à nouveau sur Continuer. La propagation de l'activation IAM peut prendre un moment.
- Aucun indicateur de progression lors de l'inscription : Sign-up cela prend quelques minutes. La page affiche une bannière statique Sign-up en cours sans barre de progression pendant qu'AWS approvisionne votre compte.
- « Connecté sous un autre compte » après avoir suivi le lien de configuration : Choisissez Déconnexion et continuez. La page vous authentifie à nouveau avec l'adresse e-mail que vous avez saisie.
- Message « Introuvable » lors de la connexion : ce message peut apparaître brièvement lors de la redirection. Tu peux le rejeter.
- La page d'utilisation n'affiche aucune donnée après votre premier appel d'API : les données d'utilisation peuvent mettre quelques minutes à apparaître dans la console Claude.

# Conditions préalables

Avant d'effectuer des appels d'API, assurez-vous d'avoir :

1. Un compte AWS actif avec un abonnement à Claude Platform sur AWS (voir Configuration de votre compte).
2. La CLI AWS est installée et configurée.
3. La fédération d'identité Web sortante est activée sur votre compte AWS (étape de configuration unique ; voir Activer la fédération d'identité Web sortante).
4. Votre identifiant d'espace de travail (voir Obtenir votre identifiant d'espace de travail).

## Activer la fédération des identités Web sortantes

La plateforme Claude sur AWS Gateway fait appel au `sts:GetWebIdentityToken` serveur pour créer un JWT qu'il transmet à Anthropic. Cette fonctionnalité STS est désactivée par défaut sur chaque compte AWS. Activez-le une fois par compte :

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

Si la réponse est oui`[ERROR] (FeatureEnabled) … already enabled` , le paramètre est déjà activé pour votre compte et vous pouvez passer à autre chose. Vérifiez et récupérez l'URL de l'émetteur de votre compte :

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

> **Warning**
>
> Sans cette étape, chaque demande est renvoyée`"Outbound web identity federation is disabled for your account"`. Il s'agit de l'erreur de configuration la plus courante.

## Obtenez votre identifiant d'espace de travail

Un espace de travail par défaut est automatiquement mis en service dans chaque région lorsque vous vous inscrivez (voir Configuration de votre compte). Les espaces de travail sont liés à une seule

région AWS. Vous trouverez l'ID de l'espace de travail dans la console Claude, sous Espaces de travail, ou dans la section Espaces de travail de la page de service de la console AWS. Consultez la section Espaces de travail pour la liaison entre les régions et l'étendue des ressources IAM.

Définissez les variables d'`AWS_REGION`environnement `ANTHROPIC_AWS_WORKSPACE_ID` et de manière à ce que les clients du SDK les lisent automatiquement :

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

La région est obligatoire. Le client du SDK lance si aucune région n'est définie. Passez`aws_region`/`awsRegion`au constructeur, ou à set `AWS_REGION` (ou`AWS_DEFAULT_REGION`). Toutes les régions commerciales AWS sont prises en charge ; pour les régions d'inscription, vous devez d'abord activer le compte dans la région.

# Authentification

Claude Platform sur AWS prend en charge deux méthodes d'authentification : AWS IAM avec signature des demandes SigV4 (principale) et authentification par clé d'API. Les deux utilisent la même URL de base et le même format de demande.

Les SDK Anthropic (voir Installer un SDK) implémentent le flux d'authentification et la résolution des identifiants décrits ci-dessous. Si vous n'utilisez pas de SDK Anthropic, vous devez créer des SigV4-signed requêtes (ou présenter votre clé d'API sous forme de jeton porteur) par rapport au point de terminaison régional. `https://aws-external-anthropic.<region>.api.aws` Pour la configuration du SDK-specific client et l'ordre de résolution des informations d'identification faisant autorité, consultez le guide de configuration de Claude on AWS dans la documentation d'Anthropic.

## Authentification SigV4

SigV4 est la solution native pour les entreprises et s'intègre à vos politiques, rôles et audits AWS IAM existants. Configurez les informations d'identification AWS à l'aide de n'importe quelle méthode prise en charge par la chaîne de fournisseurs d'informations d'identification par défaut d'AWS :

- Variables d'environnement (`AWS_ACCESS_KEY_ID`,`AWS_SECRET_ACCESS_KEY`,`AWS_SESSION_TOKEN`)
- Fichier d'informations d'identification partagé (`~/.aws/credentials`)
- Fichier de configuration partagé (`~/.aws/config`), y compris SSO et `credential_process`
- Identité Web (`AWS_WEB_IDENTITY_TOKEN_FILE`et`AWS_ROLE_ARN`) pour IRSA et Actions GitHub
- Informations d'identification du conteneur ECS
- Service de métadonnées d'instance EC2 (IMDS)

Pour vérifier que le SDK Anthropic récupère vos informations d'identification et les résout vers le point de terminaison régional approprié, effectuez une demande de test en suivant les exemples de la section Faire des demandes. Une réponse positive confirme que les informations d'identification, la région, l'URL de base et l'ID de l'espace de travail sont tous correctement configurés.

## Authentification par clé d'API

Pour des chemins d'intégration plus simples (développement local et scripts), vous pouvez vous authentifier avec une clé API au lieu de Sigv4. Définissez la variable

d'`ANTHROPIC_AWS_API_KEY`environnement ou `apiKey` passez-la au constructeur du SDK. Le SDK Anthropic envoie la clé sous forme de jeton porteur au point de terminaison Claude Platform on AWS ; IAM autorise la demande par le biais de l'`aws-external-anthropic:CallWithBearerToken`action (voir les politiques IAM).

Générez des clés d'API dans la console AWS sous Claude Platform sur AWS → Clés d'API. Choisissez Générer une clé, puis copiez la valeur de la clé. Accordez l'action `aws-external-anthropic:CallWithBearerToken` IAM aux principaux qui devraient être autorisés à utiliser l'authentification par clé d'API.

> **Note**
>
> Seules les clés d'API créées dans la console AWS sous la plateforme Claude sur AWS fonctionnent avec ce service.
>
> - Les clés créées dans la console Claude standard pour l'accès à l'API de première partie ne fonctionnent pas avec la plateforme Claude sur le point de terminaison AWS.
> - Les clés d'API Amazon Bedrock ne fonctionnent pas non plus : Bedrock utilise un point de terminaison, un flux d'authentification et un espace de noms IAM distincts.

## Short-term Clés d'API

Dans les cas où vous souhaitez une authentification par clé d'API mais sans la durée de vie d'une clé de longue durée, Anthropic publie des bibliothèques de générateurs de jetons qui créent des clés d'API à court terme à partir de vos informations d'identification AWS IAM. Les jetons ont par défaut une durée de vie de 12 heures et peuvent être étendus à un espace de travail spécifique et à l'`aws-external-anthropic:CallWithBearerToken`action. Installez le générateur correspondant à votre langue, échangez les informations d'identification IAM contre une clé d'API et transmettez la clé au SDK Anthropic.

- Python : aws/token-generator-for-aws-external-anthropic-python
- JavaScript / TypeScript: aws/token-generator-for-aws-external-anthropic-js
- Java : aws/token-generator-for-aws-external-anthropic-java

Short-term Les clés d'API nécessitent toujours que le principal IAM de l'appelant conserve l'espace `aws-external-anthropic:CallWithBearerToken` de travail cible. Ils expirent d'eux-mêmes et n'ont pas besoin d'être changés ou révoqués explicitement.

## Priorité des informations d'identification et résolution des régions

La plateforme Claude du SDK Anthropic sur le client AWS résout les informations d'identification et la région en utilisant un ordre de priorité défini. Les noms des arguments suivent les conventions de chaque langage : CamelCase TypeScript pour et PHP, snake_case pour Python et Ruby PascalCase , pour Go, et modèles de propriété ou de générateur pour C# et Java.

Pour connaître l'ordre de priorité officiel, les variables d'environnement prises en charge et les arguments du constructeur pour chaque SDK, consultez la section Configuration de Claude on AWS dans la documentation d'Anthropic. En général, les arguments explicites du constructeur ont priorité sur les variables d'environnement, ainsi que sur la `ANTHROPIC_AWS_API_KEY` chaîne de fournisseurs d'informations d'identification AWS par défaut. La région est obligatoire ; contrairement à `AnthropicBedrock` (qui revient à`us-east-1`), le client Claude on AWS renvoie si aucune région n'est fournie par le constructeur ou par`AWS_REGION`/`AWS_DEFAULT_REGION`.

## ID d'espace de travail

Chaque demande de plan de données doit inclure l'ID de l'espace de travail dans l'`anthropic-workspace-id`en-tête. Les SDK Anthropic lisent cela à partir de la variable d'`ANTHROPIC_AWS_WORKSPACE_ID`environnement par défaut, ou vous pouvez transmettre`workspaceId`/`workspace_id`au constructeur du client. Si vous utilisez le `Anthropic` client de base (et non le Claude-on-AWS client), transmettez explicitement l'en-tête. Consultez les sections Faire des demandes d'exemples par langue et Espaces de travail pour savoir comment localiser votre identifiant d'espace de travail.

# Installer un kit SDK

Les SDK clients d'Anthropic prennent en charge la plateforme Claude sur AWS. Les sept SDK fournissent une classe de client spécifique à la plate-forme ; vous pouvez également configurer le client de base avec l'URL de base de la plateforme Claude sur AWS.

Python

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

> **Tip**
>
> Sur macOS avec Homebrew Python ou d'autres environnements Python gérés en externe, cela `pip install` peut échouer avec une erreur PEP `externally-managed-environment` 668. Créez et activez d'abord un environnement virtuel :`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**
>
> Les clients du SDK pour Claude Platform sur AWS sont en version bêta.

# Modèles disponibles

La plateforme Claude sur AWS propose le même ensemble de modèles Claude que l'API Claude propriétaire.

Pour la liste actuelle des modèles, des identifiants de modèles, des tailles de fenêtres contextuelles et des fonctionnalités, consultez la section Présentation des modèles dans la documentation d'Anthropic.

## ID des modèles

Les identifiants de modèle sur la plateforme Claude sur AWS sont identiques à ceux utilisés sur l'API Claude propriétaire (par exemple, `claude-opus-4-6`claude-sonnet-4-6,`claude-haiku-4-5`). Il n'y a pas Bedrock-style d'ARN ni de `anthropic.` préfixe. Utilisez l'identifiant du modèle exactement tel qu'Anthropic le publie.

# Demandes

La plateforme Claude sur AWS utilise l'API Anthropic Messages (`/v1/messages`), la même surface d'API que la plateforme propriétaire Claude. Les différences concernent l'URL de base, la méthode d'authentification et un `anthropic-workspace-id` en-tête obligatoire qui identifie l'espace de travail cible par la demande.

Coquille

```
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
```

Le client lit `AWS_REGION` (ou`AWS_DEFAULT_REGION`) et `ANTHROPIC_AWS_WORKSPACE_ID` depuis l'environnement. Vous pouvez le remplacer en passant`aws_region`/`awsRegion`ou`workspace_id`/`workspaceId`au constructeur. La région et l'identifiant de l'espace de travail sont requis ; le constructeur les renvoie si l'un ou l'autre ne peut pas être résolu à partir de ces sources.

> **Note**
>
> L'`x-amz-security-token`en-tête (curl) n'est requis que pour les informations d'identification temporaires telles que les rôles IAM, SSO ou STS. Omettez-le lorsque vous utilisez des informations d'identification utilisateur IAM à long terme. Les clients du SDK gèrent cela automatiquement en fonction de la source des informations d'identification.

La `--aws-sigv4` valeur suit le format`aws:amz:<region>:<service>`. Le nom du service Sigv4 est`aws-external-anthropic`, et la région doit correspondre à la région dans l'URL de votre point de terminaison. Une incompatibilité entre l'un ou l'autre produit une erreur générique de rejet de signature plutôt qu'un diagnostic spécifique.

## Utilisation du client Anthropic de base

Si vous préférez ne pas ajouter de dépendance au SDK AWS, vous pouvez utiliser le `Anthropic` client de base. Ce chemin utilise les noms des variables d'environnement du SDK vanilla plutôt que les `ANTHROPIC_AWS_*` noms lus par le client dédié :

```
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'
```

Les SDK Python et Go lisent `ANTHROPIC_API_KEY` et se lisent `ANTHROPIC_BASE_URL` automatiquement ; pour les autres langages, transmettez-les au constructeur. TypeScript Dans chaque langue, transmettez l'identifiant de l'espace de travail dans l'`anthropic-workspace-id`en-tête.

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);
```

# Prise en charge des fonctionnalités

Claude Platform sur AWS utilise directement l'API Anthropic Messages. Vous bénéficiez de la parité complète des fonctionnalités avec l'API Claude propriétaire, sauf indication contraire dans la section Fonctionnalités non disponibles actuellement :

- Fonctionnalités bêta : transmettez l'`anthropic-beta`en-tête standard pour accéder aux fonctionnalités bêta, comme vous le feriez avec l'API Claude.
- Compétences d'agent : utilisez des compétences d'agent prédéfinies et personnalisées avec les mêmes `container.skills` paramètres et en-têtes bêta que l'API Claude. Toutes les compétences PowerPoint prédéfinies (Excel, Word, PDF) fonctionnent immédiatement.
- Exécution du code : exécutez le code dans le sandbox géré d'Anthropic à l'aide de l'outil d'exécution de code.
- Utilisation des outils : L'utilisation de l'ordinateur et toutes les autres fonctionnalités d'utilisation des outils sont disponibles.
- Pensée étendue : activez la réflexion étendue avec les mêmes paramètres que l'API Claude.
- Streaming : support complet du streaming SSE pour des réponses en temps réel.
- Traitement par lots : soumettez des demandes par lots pour les charges de travail à haut débit.
- Mise en cache rapide : outils de mise en cache, instructions du système et historique des messages pour réduire la latence et les coûts. Toutes les fonctionnalités de mise en cache rapide (5 minutes TTL, 1 heure TTL et mise en cache automatique) sont disponibles.
- API de fichiers : téléchargez et référencez des fichiers dans toutes les demandes.
- Balises d'espace de travail avec intégration IAM : les espaces de travail peuvent être balisés, et les balises sont transmises à IAM pour le contrôle d'accès basé sur les attributs et aux rapports sur les coûts et l'utilisation d'AWS pour la répartition des coûts. Le balisage de l'espace de travail est une fonctionnalité GA de la plateforme Claude sur AWS (il s'agit d'une fonctionnalité bêta de l'API Claude propriétaire).

# Agents gérés par Claude

Les agents gérés par Claude sont disponibles sur la plateforme Claude sur AWS. Les agents, les sessions, les environnements, les coffres-forts d'informations d'identification et les magasins de mémoire sont des ressources IAM de premier ordre. Voir Actions IAM pour la référence des actions et Utilisation de la console Claude pour les pages correspondantes. Le SDK d'Anthropic Agent

fonctionne avec Claude Platform sur AWS sans intégration supplémentaire. Dirigez-le vers l'URL de base de la plateforme Claude sur AWS et authentifiez-vous avec SigV4 ou une clé d'API.

Les sessions autonomes sur la plateforme Claude sur AWS nécessitent une nouvelle authentification toutes les 6 heures. Long-running les exécutants de l'agent doivent actualiser leurs informations d'identification Sigv4 ou leur clé API dans cette fenêtre, sinon la session se termine.

Les fonctionnalités suivantes des agents gérés par Claude ne sont actuellement pas disponibles sur la plateforme Claude sur AWS :

- Résultats : Le suivi des résultats pour les sessions avec les agents n'est pas disponible.
- Multi-agent sessions : les sessions avec plusieurs agents en interaction ne sont pas disponibles.

## Conformité d'

> **Important**
>
> Anthropic fournit ce service en tant qu'offre tierce. Il n'est pas couvert par les programmes de conformité, les certifications ou les rapports d'audit standard d'AWS (tels que l'éligibilité aux normes SOC, ISO ou HIPAA). Les clients sont seuls responsables de faire preuve de diligence raisonnable pour s'assurer que cette offre tierce répond à leurs exigences réglementaires, légales et de conformité. Avant de traiter des données sensibles ou réglementées, vous devez consulter la documentation de conformité officielle d'Anthropic via le Anthropic Trust Center. Consultez les conditions de service AWS pour plus d'informations.

## Comment est modélisée l'adhésion à l'espace de travail

Sur l'API Claude propriétaire, l'accès est régi par l'adhésion des utilisateurs aux espaces de travail : un utilisateur est ajouté à un espace de travail et hérite des droits d'accès de l'espace de travail. Claude Platform sur AWS remplace ce modèle par une autorisation IAM : il n'existe aucune liste d'utilisateurs par espace de travail. Au lieu de cela, les principaux IAM (utilisateurs ou rôles) appliquent des politiques qui autorisent des `aws-external-anthropic` actions contre des ARN spécifiques de l'espace de travail. L'évaluation de la politique IAM détermine ce que chaque directeur peut faire dans chaque espace de travail.

Accordez et révoquez l'accès à l'espace de travail en modifiant les politiques IAM (SCP, limites d'autorisation, politiques liées à l'identité ou conditions au niveau des ressources) plutôt que de

gérer l'adhésion par espace de travail dans la console Claude. Consultez les politiques IAM pour des exemples.

## Fonctionnalités non disponibles actuellement

Les fonctionnalités suivantes ne sont actuellement pas disponibles sur la plateforme Claude sur AWS :

- Préparation à la loi HIPAA : le HIPAA-ready programme d'Anthropic n'est pas disponible sur la plateforme Claude sur AWS. Les clients soumis aux exigences de la loi HIPAA devraient plutôt évaluer Claude dans Amazon Bedrock.
- Niveaux de service autres que Standard et Batch : le niveau prioritaire n'est pas disponible.
- API d'administration : membres de l'organisation, membres de l'espace de travail, invitation, clé d'API, rapport d'utilisation, rapport de coûts et points de terminaison du rapport sur les limites de débit : ces points de terminaison de l'API d'administration ne sont pas disponibles actuellement. Workspace CRUD et rename endpoints (`/v1/organizations/workspaces`) sont disponibles via l'espace de `aws-external-anthropic` noms ; voir Actions IAM. L'adhésion est modélisée par le biais d'AWS IAM plutôt que par le biais des listes de membres des espaces de travail (voir la section précédente). Le cycle de vie des clés d'API est géré dans la console AWS sous Claude Platform sur AWS → Clés d'API. Pour les données d'utilisation, de coût et de limite de débit, utilisez les vues de la console Claude (voir Surveillance et journalisation) ou l'API Anthropic Usage and Cost.
- Limites de dépenses : Non disponible. Fiez-vous plutôt aux contrôles de facturation d'AWS.
- Espace de travail Claude Code et API Analytics : L'espace de travail Claude Code avec limites de débit automatiques n'est pas disponible. Claude L'utilisation du code apparaît dans la vue d'utilisation générale plutôt que dans un écran dédié.
- Authentification OAuth : non prise en charge. Utilisez l'authentification par SigV4 ou par clé API.
- OpenAI-compatible Points de terminaison d'API : non disponibles sur la plateforme Claude sur AWS.
- Workspace-level contrôles géographiques d'inférence : `allowed_inference_geos` et ne `default_inference_geo` sont pas disponibles. Définissez plutôt `inference_geo` sur chaque demande.

# Résidence des données

La plateforme Claude sur AWS prend en charge les zones d'inférence suivantes :

- États-Unis : l'inférence reste dans les centres de données américains. Un multiplicateur de prix de 1,1x s'applique.
- Global : L'inférence peut être acheminée vers n'importe quel centre de Anthropic-operated données dans le monde entier. La tarification standard s'applique.

Définissez la géographie d'inférence par demande avec le `inference_geo` paramètre :

> **Note**
>
> Le `inference_geo` paramètre est pris en charge sur les modèles Claude Opus 4.6, Claude Sonnet 4.6 et les modèles ultérieurs. Les requêtes utilisant `inference_geo` Claude Opus 4.5, Claude Sonnet 4.5 ou Claude Haiku 4.5 renvoient une erreur 400. Voir Résidence des données pour plus de détails sur la disponibilité des modèles.

Coquille

```
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 vous omettez`inference_geo`, la valeur par défaut de la demande est. `global`

Workspace-level les contrôles géographiques d'inférence (`allowed_inference_geos`et`default_inference_geo`) ne sont pas disponibles sur la plateforme Claude sur AWS. Définissez plutôt `inference_geo` sur chaque demande.

# Espaces de travail

Les demandes d'inférence et de ressources sur Claude Platform sur AWS ciblent un espace de travail. Vous transmettez l'ID de l'espace de travail dans l'`anthropic-workspace-id`en-tête de ces appels d'API. Les identifiants d'espace de travail utilisent le format balisé `wrkspc_` suivi d'un identifiant alphanumérique (par exemple,`wrkspc_01AbCdEf23GhIj`). Consultez Obtenir votre identifiant d'espace de travail si vous ne l'avez pas encore.

## Délimitation de l'espace de travail

Les espaces de travail sont liés à une seule région AWS. Un espace de travail créé dans n'est `us-west-2` accessible que via le `us-west-2` point de terminaison. L'utilisation, les quotas, les coûts, les fichiers, les lots et les compétences sont tous cumulés par espace de travail, vous donnant des ventilations par région dans la console Claude.

Les espaces de travail constituent également la principale ressource IAM pour Claude Platform sur AWS. Vous accordez ou refusez l'accès à des espaces de travail spécifiques par le biais des politiques AWS IAM à l'aide de l'ARN de l'espace de travail. Le segment de ressource de l'ARN est le même identifiant `wrkspc_` préfixé que vous transmettez dans l'`anthropic-workspace-id`en-tête :

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

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

Consultez les politiques IAM pour des exemples de politiques.

## Création d'espaces de travail

Un espace de travail par défaut est mis en service automatiquement lors de l'inscription. Des espaces de travail supplémentaires peuvent être créés, mis à jour, renommés et archivés via les `/v1/organizations/workspaces` points de terminaison, autorisés par les `aws-external-anthropic` actions correspondantes (`CreateWorkspace`,`UpdateWorkspace`,`ArchiveWorkspace`) ; voir Actions IAM.

# Configuration de l'identifiant de l'espace de travail dans votre client

Chaque demande de plan de données doit inclure l'ID de l'espace de travail dans l'`anthropic-workspace-id`en-tête. Lorsque vous utilisez le Claude-on-AWS client d'un SDK Anthropic, il existe trois manières de le fournir :

1. Variable d'environnement : définie`ANTHROPIC_AWS_WORKSPACE_ID`, et le client la lit automatiquement.

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

2. Argument du constructeur : passez`workspaceId`/`workspace_id`(le nom suit la convention linguistique du SDK) lors de l'instanciation du client.
3. Per-request override : transmettez l'en-tête directement sur une demande individuelle si vous devez adresser plusieurs espaces de travail à partir du même client.

Si vous utilisez le `Anthropic` client de base (sans le backend AWS), définissez l'`anthropic-workspace-id`en-tête de manière explicite pour chaque demande. Consultez la section Faire des demandes pour des exemples par langue. Pour les noms des SDK-specific arguments, reportez-vous à la documentation des SDK d'Anthropic Client.

# Identification des espaces de travail

Les balises d'espace de travail sont des paires clé-valeur associées à un espace de travail. Ils s'intègrent à AWS IAM pour le contrôle d'accès basé sur les attributs, et aux rapports sur les coûts et l'utilisation d'AWS pour la répartition des coûts (voir Surveillance et journalisation pour plus de détails sur le CUR). Le balisage s'effectue en GA sur Claude Platform sur AWS.

Mettez à jour les balises de l'espace de travail existant avec `TagResource` et `UntagResource` dans l'espace de `aws-external-anthropic` noms. Les mêmes clés de balise peuvent définir les conditions IAM et la répartition des coûts sans configuration supplémentaire.

## Tag-based contrôle d'accès

Vous pouvez bloquer `aws-external-anthropic` des actions sur les valeurs des balises de l'espace de travail à l'aide de la clé contextuelle de `aws:ResourceTag/<key>` condition. La politique suivante autorise l'inférence uniquement sur les espaces de travail balisés : `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"
        }
      }
    }
  ]
}
```

Utilisez `aws:RequestTag/<key>` et `aws:TagKeys` conditionnez les touches activées `TagResource` pour appliquer les politiques de balisage (par exemple, exiger que l'espace de travail comporte `CostCenter` des `Owner` étiquettes). Voir les politiques IAM pour d'autres exemples.

# Utilisation de la console Claude

La plateforme Claude sur AWS utilise la console Claude standard disponible sur platform.claude.com. Lorsque vous vous connectez depuis la console AWS, un indicateur de compte géré par AWS apparaît dans la barre latérale de la console Claude. La console s'applique à votre organisation Claude Platform on AWS. Il fournit des analyses d'utilisation, une ventilation des coûts, une visibilité des limites de débit, une visibilité de l'espace de travail et des pages pour la gestion des fichiers, des compétences des agents, des tâches par lots, des agents, des sessions, des environnements, des coffres-forts d'informations d'identification et des stockages de mémoire.

## Connexion

L'accès à la console Claude est fédéré via AWS IAM. Consultez la section Configurer votre compte pour connaître le flux complet de connexion pour la première fois. En bref :

1. Assumez un rôle IAM avec l'`aws-external-anthropic:AssumeConsole`autorisation. Consultez la section Actions IAM.
2. Accédez à la page Claude Platform on AWS dans la console AWS.
3. Choisissez Open Claude Console. La console AWS émet un JWT et vous redirige vers. `platform.claude.com`
4. Lors de votre première connexion, vous êtes invité à saisir une adresse e-mail ; entrez votre adresse e-mail professionnelle. La plateforme approvisionne votre utilisateur de la console Claude juste à temps.

Deux rôles Claude Console sont disponibles : administrateur et développeur. Le rôle d'administrateur donne accès à toutes les pages et à tous les paramètres de la console Claude disponibles pour Claude Platform sur AWS. Le rôle de développeur accorde un accès en lecture aux informations d'utilisation, de coût, de limite de débit et d'espace de travail. Contactez le représentant de votre compte Anthropic pour attribuer le rôle d'administrateur ou de développeur à un mandant.

## Pages disponibles

La colonne Via AWS Gateway indique si la page lit et écrit des données via la passerelle AWS (et est donc régie par les actions IAM). Les pages marquées Non lisent les métadonnées au niveau de l'organisation directement via les API Anthropic et contournent les contrôles d'action IAM.

| Page | Available | Par le biais de la passerelle AWS | Remarques |
|---|---|---|---|
| Utilisation | Oui | Non | Affichez l'utilisation des jetons par modèle, espace de travail et dimension. Les données peuvent prendre quelques minutes avant d'apparaître après une demande. |
| Coût | Oui | Non | Consultez la répartition des coûts par modèle et par espace de travail. AWS Cost Explorer affiche l'élément de ligne CCU agrégé. |
| Restrictions | Oui | Non | Afficher les limites de débit (lecture seule). |
| Espaces de travail | Oui | Non | Afficher les espaces de travail par région (lecture seule). |
| Dépôt de | Oui | Oui | Afficher et gérer les fichiers téléchargés. |
| Compétences | Oui | Oui | Consultez et gérez les compétences des agents. |
| Lots | Oui | Oui | Affichez et gérez les tâches de traitement par lots. |
| Agents | Oui | Oui | Affichez et gérez les définitions des agents. |
| Sessions | Oui | Oui | Consultez les sessions des agents et l'histori que des événements. |
| Environne ments | Oui | Oui | Affichez et gérez les configurations de conteneurs cloud pour les sessions. |
| Coffres-forts d'informa tions d'identif ication | Oui | Oui | Affichez et gérez les coffres d'informations d'identification pour l'authentification de session. |

| Page | Available | Par le biais de la passerelle AWS | Remarques |
|---|---|---|---|
| Magasins de mémoire | Oui | Oui | Affichez et gérez la mémoire persistante des agents. |
| Clés d'API | Non | N/A | Gérez les clés d'API dans la console AWS (Claude Platform sur AWS → clés d'API). Voir Authentification. |
| Membres | Non | N/A | Non applicable. AWS IAM gère les accès. |
| Facturation | Non | N/A | Non applicable. AWS Marketplace gère la facturation et la facturation. Consultez la ventilation des coûts sur la page des coûts. |
| Claude Code | Non | N/A | Consultez l'utilisation du code Claude sur la page Utilisation. |

## Changer d'organisation

La console Claude ne prend pas en charge le changement d'organisation pour Claude Platform sur AWS. Pour accéder à une autre organisation, déconnectez-vous et authentifiez-vous à nouveau via la console AWS en utilisant le rôle IAM pour le compte AWS de cette organisation.

# Limites de taux et quotas

Claude Platform sur AWS attribue des limites de débit de niveau 1 lorsque vous vous abonnez. Anthropic gère les limites de débit directement, et non par le biais des systèmes de quotas AWS.

## Limites par défaut

Claude Platform sur AWS utilise le barème de niveaux standard d'Anthropic, identique à celui de l'API Claude propriétaire. Les limites de niveau 1 s'appliquent par espace de travail. Les limites sont regroupées par famille de modèles. Par exemple, tous les modèles Opus partagent une limite combinée, tous les modèles Sonnet en partagent une autre et tous les modèles Haiku en partagent une troisième.

Pour les valeurs de niveau 1 actuelles (RPM, ITPM, OTPM) et les seuils de niveau supérieur, voir Limites de débit sur le site Web de documentation d'Anthropic. La page Anthropic est la source de vérité et est mise à jour lorsque les limites changent.

## En-têtes de limite de débit

Chaque réponse inclut des en-têtes qui indiquent l'état actuel de votre limite de débit. En-têtes principaux :

- `anthropic-ratelimit-requests-limit`— Nombre maximum de demandes par minute
- `anthropic-ratelimit-requests-remaining`— Demandes restantes dans la fenêtre actuelle
- `anthropic-ratelimit-requests-reset`— Heure à laquelle la limite de demandes est réinitialisée (RFC 3339)
- `anthropic-ratelimit-tokens-limit`— Nombre maximal de jetons combinés (entrée+sortie) par minute
- `anthropic-ratelimit-tokens-remaining`— Jetons combinés restant dans la fenêtre actuelle
- `anthropic-ratelimit-tokens-reset`— Heure à laquelle la limite de jetons combinés est réinitialisée (RFC 3339)
- `anthropic-ratelimit-input-tokens-limit/-remaining/-reset`— Input-token-specific en-têtes

- `anthropic-ratelimit-output-tokens-limit/-remaining/-reset`— Output-token-specific en-têtes
- `retry-after`— Sur une réponse 429, le nombre de secondes à attendre avant de réessayer

Voir les en-têtes de réponse sur le site Web de documentation d'Anthropic pour l'ensemble complet.

## Demande de limites plus élevées

Contrairement à l'API Claude propriétaire, l'avancement automatique des niveaux ne s'applique pas à la plateforme Claude sur AWS. Pour demander des limites plus élevées, contactez le représentant de votre compte Anthropic avec votre identifiant d'espace de travail et le débit souhaité. Pour les seuils de niveau et d'autres détails, consultez les limites de taux sur le site Web de documentation d'Anthropic.

## Erreurs de limite de débit

Lorsque vous dépassez une limite de débit, l'API renvoie le protocole HTTP 429 avec un `rate_limit_error` type. Implémentez un ralentissement exponentiel avec instabilité dans votre logique de nouvelle tentative. L'`retry-after`en-tête indique le nombre de secondes à attendre avant de réessayer.

# Facturation

Claude Platform sur AWS utilise AWS Marketplace comme backend de facturation pour le comptage basé sur la consommation. Anthropic mesure l'utilisation horaire des compteurs via AWS Marketplace, et AWS émet des factures mensuelles sur votre facture AWS existante.

La facturation s'effectue uniquement pour les arriérés (vous payez pour ce que vous avez utilisé) et avant impôts (AWS applique les paramètres fiscaux de votre compte).

L'utilisation est exprimée en unités de consommation Claude (UCC) à 0,01 USD par UCC. Le prix du CCU est fixe et n'est jamais réduit. Anthropic évalue l'utilisation de vos jetons en USD aux taux standard par modèle et par fonctionnalité, applique toute réduction négociée, puis convertit le résultat en CCU à 0,01$ par CCU. Les remises se traduisent par une diminution du nombre de CCU mesurées, et non par une baisse du prix des CCU. Les CCU ne sont pas des crédits prépayés ; il n'y a aucun solde ou engagement de CCU. Consultez la section Tarification pour la définition de la CCU et les taux de jetons par modèle.

# Surveillance et journalisation

AWS CloudTrail peut capturer toutes les demandes adressées à Claude Platform sur AWS. Les opérations de l'espace de travail et du coffre-fort sont enregistrées en tant qu'événements de gestion par défaut. Toutes les autres opérations (inférence, lot, fichier, compétence, modèle, profil utilisateur et agents gérés par Claude, à l'exception des coffres-forts) sont des événements de données. Leur enregistrement nécessite une configuration explicite et entraîne des CloudTrail frais supplémentaires. Consultez les actions IAM pour la classification complète des types d'événements et la CloudTrail documentation AWS pour les détails de configuration.

Lors de la configuration CloudTrail de la journalisation des événements de données, sélectionnez le type de ressource`AWS::AWSExternalAnthropic::Workspace`. Il s'agit du type de CloudTrail ressource pour Claude Platform sur les espaces de travail AWS. Il est nécessaire pour capturer les événements du plan de données (inférence, traitement par lots, fichiers et autres opérations par espace de travail). Étendez les sélecteurs d'événements de données aux ARN spécifiques de l'espace de travail si vous souhaitez enregistrer l'activité d'un sous-ensemble d'espaces de travail plutôt que de l'ensemble du compte.

## Identifiants de demande

Chaque réponse inclut deux identifiants de demande dans les en-têtes de réponse :

- ID de demande AWS (**`x-amzn-requestid`**) : ID principal, indexé dans CloudTrail. Utilisez-le lorsque vous étudiez des demandes via les outils AWS ou lorsque vous contactez le support AWS.
- ID de demande anthropique (**`request-id`**) : ID secondaire. Utilisez-le lorsque vous contactez l'assistance d'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

Le SDK Ruby n'expose actuellement aucun accesseur à réponse brute. Les identifiants de demande ne peuvent donc pas être lus à partir des en-têtes de réponse dans Ruby. Si vous avez besoin de la journalisation des identifiants de demande, utilisez l'une des autres langues ci-dessus ou capturez les identifiants au niveau de la couche proxy HTTP.

Anthropic recommande d'enregistrer votre activité sur une base continue d'au moins 30 jours pour comprendre les habitudes d'utilisation et étudier tout problème potentiel.

> **Note**
>
> AWS CloudTrail est configuré dans votre compte AWS. L'activation de la journalisation ne permet pas à AWS ou à Anthropic d'accéder à votre contenu au-delà de ce qui est nécessaire pour la facturation et le fonctionnement du service.

## Indicateurs d'utilisation et tableaux de bord

Claude Platform sur AWS ne publie pas de statistiques d'utilisation par espace de travail sur Amazon CloudWatch. Le nombre de jetons, le volume de demandes et l'utilisation par modèle sont plutôt disponibles sur les surfaces d'utilisation d'Anthropic :

- Vues d'utilisation de la console Claude : lorsqu'un directeur se fédère dans la console Claude avec `aws-external-anthropic:AssumeConsole` (voir les politiques IAM), les tableaux de bord d'utilisation indiquent le volume de demandes, le nombre de jetons et les ventilations par modèle pour les espaces de travail accessibles. Account-wide les vues d'utilisation nécessitent la fonctionnalité de console d'administration.
- API d'utilisation et de coût anthropiques : pour un accès programmatique aux données d'utilisation et de coûts, y compris l'agrégation par espace de travail et par modèle, voir API d'utilisation et de coût dans la documentation d'Anthropic.

Pour la surveillance opérationnelle (taux d'erreur, latence, en-têtes de limite de débit), utilisez les en-têtes de réponse renvoyés à chaque demande (voir Limites de débit et quotas) ainsi que les identifiants de demande AWS capturés dans. CloudTrail

## Allocation et suivi des coûts

Les frais liés à la plateforme Claude sur AWS apparaissent sur votre facture AWS dans le cadre du service Claude Platform on AWS, avec les dimensions d'utilisation par modèle. Utilisez le rapport sur les coûts et l'utilisation (CUR) d'AWS associé aux balises d'espace de travail pour une répartition précise des coûts :

1. Marquez vos espaces de travail avec les dimensions d'allocation qui vous intéressent (équipe, application, environnement, centre de coûts). Consultez la section Espaces de travail pour plus de détails sur le balisage.

2. Dans la console de facturation AWS, activez chaque clé de balise en tant que balise de répartition des coûts. Après l'activation, l'utilisation effectuée à la date d'activation ou après cette date est divisée par étiquette dans le CUR.
3. Dans CUR ou AWS Cost Explorer, regroupez-les par `resourceTags/user:<tag-key>` colonne pour ventiler les dépenses par balise d'espace de travail.

Les balises de l'espace de travail sont transmises aux éléments de ligne CUR pour toutes les utilisations de la plateforme Claude sur AWS. Les mêmes clés de balise régissent à la fois le contrôle IAM-based d'accès et la répartition des coûts. Consultez les balises de répartition des coûts dans la documentation de facturation AWS pour connaître les étapes de configuration et les délais d'activation.

# Migrer depuis Amazon Bedrock

Si vous utilisez actuellement Claude sur Bedrock, la migration vers Claude Platform sur AWS nécessite des modifications tout au long de votre intégration. La signature SigV4 reste prise en charge, mais le contexte de signature, l'URL de base, le format d'API, les ID de modèle, le client et le package du SDK, le format de streaming, les en-têtes de requête et la disponibilité des régions changent tous. Le tableau suivant récapitule les différences.

## Quels sont les changements

| | Amazon Bedrock | Claude Platform sur AWS |
|---|---|---|
| URL de base | `bedrock-runtime.{region}.amazonaws.com` | `aws-external-anthropic.{region}.api.aws` |
| Format de l'API | Bedrock Converse/ InvokeModel | API de messages anthropiques () `/v1/messages` |
| Identifiants du modèle | `anthropic.claude-opus-4-7-v1` | `claude-opus-4-7` |
| Client du SDK | `AnthropicBedrock` /Kit de développement logiciel Bedrock | Platform-specific client (`AnthropicAWS` ,`AnthropicAws` ,`AnthropicAwsClient` , etc.), en version bêta |
| Package SDK | `anthropic[bedrock]` `@anthropic-ai/bedrock-sdk`, etc. | `anthropic[aws]` `@anthropic-ai/aws-sdk` , etc. (voir Installer un SDK) |
| Nom du service SigV4 | `bedrock` | `aws-external-anthropic` |
| Format de diffusion | AWS EventStream | SSE (identique à l'API Claude) |
| En-tête d'espace | Non applicable | `anthropic-workspace-id` obligatoire |

| | Amazon Bedrock | Claude Platform sur AWS |
|---|---|---|
| Disponibilité dans les Régions | Plusieurs régions commerciales | Toutes les régions commercia les AWS (les régions optionnelles nécessitent l'inscription d'un compte) |

## Ce que vous gagnez

- Accès généralement le jour même aux nouveaux modèles et fonctionnalités, sans étape d'intégration séparée avec les partenaires
- Compétences des agents pour la génération de documents (ExcelPowerPoint, Word, PDF)
- Exécution du code dans le sandbox géré d'Anthropic
- Fonctionnalités bêta via l'`anthropic-beta`en-tête (voir Fonctionnalités non disponibles actuellement)
- Console Claude pour la visibilité des quotas et l'analyse de l'utilisation
- Soutien anthropique direct
- Authentification par clé API comme alternative à SigV4 (voir Authentification)

## Ce qui reste le même

- Authentification AWS IAM (SigV4)
- Facturation via votre compte AWS
- Engagement AWS : départ à la retraite

## Les pièges de la migration

> **Warning**
>
> Activez d'abord la fédération des identités Web sortantes. Si votre compte AWS n'a jamais utilisé la plateforme Claude sur AWS, vous devez activer la fédération d'identité Web sortante une fois par compte avant de faire des demandes. Sans cette étape, toutes les demandes échouent avec une erreur de fédération. Cette étape n'est pas obligatoire pour Bedrock.

> **Warning**
>
> La conservation zéro des données (ZDR) est facultative sur la plateforme Claude sur AWS. Sur Bedrock, AWS est le processeur des données et Anthropic ne conserve ni les entrées ni les sorties d'inférence ; le programme ZDR d'Anthropic ne s'y applique pas. Sur Claude Platform sur AWS, Anthropic est le processeur des données, et ZDR suit le modèle d'API Claude propriétaire : il est disponible sur demande auprès du représentant de votre compte Anthropic. Confirmez l'inscription au ZDR avant de migrer les charges de travail de production qui dépendent des garanties de conservation des données.

# Considérations commerciales

- Conditions d'utilisation : Les conditions de service AWS régissent la plateforme Claude sur AWS. Les conditions d'utilisation commerciales, l'addendum relatif au traitement des données et la politique d'utilisation d'Anthropic s'appliquent également.
- Réductions et offres privées : les remises négociées et les offres privées AWS Marketplace ne sont pas transférées automatiquement entre Bedrock et Claude Platform sur AWS. Collaborez avec le représentant de votre compte Anthropic pour définir les conditions commerciales de la plateforme Claude sur AWS.

# politiques IAM

Claude Platform sur AWS s'intègre à AWS IAM pour le contrôle d'accès. Vous accordez ou refusez l'accès à des actions d'API spécifiques sur des espaces de travail spécifiques en utilisant la syntaxe de politique IAM standard.

Le nom du service SigV4 et l'espace de noms des actions IAM sont. `aws-external-anthropic` Les actions suivent le modèle `aws-external-anthropic:<Action>` (par exemple,`aws-external-anthropic:CreateInference`).

## Exemple : refuser l'inférence par lots

La politique suivante permet une inférence en temps réel tout en bloquant le traitement par lots, une exigence courante pour les charges ZDR-sensitive de travail :

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

```
}
```

L'`GetBatchInference`action autorise à la fois la route des métadonnées par lots et la route des résultats par lots. Le refuser bloque également `ListBatchInferences` les lectures et l'énumération par lots.

L'`Allow`instruction énumère des `List*` actions `Get*` et des actions spécifiques plutôt que d'utiliser des caractères génériques. Les caractères génériques peuvent autoriser `GetFile` (ce qui télécharge les octets du fichier) et d'autres lectures que vous n'avez peut-être pas l'intention d'effectuer ; des `Deny` remplacements sont `Allow` malgré tout possibles, mais la forme explicite est le modèle le plus sûr à modéliser.

## Exemple : inférence synchrone sur un seul espace de travail

Accorde les autorisations minimales à un principal IAM qui exécute une inférence sur un espace de travail de production :

```
{
  "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**
>
> Le `List*` caractère générique de cette politique correspond également`ListWorkspaces`, ce qui est limité au compte. La contrainte ARN de l'espace de travail le filtre silencieusement. Cette politique n'autorise donc pas la mise en liste des espaces de travail. Si votre compte

de service doit énumérer les espaces de travail, ajoutez une `Allow` instruction distincte pour `ListWorkspaces` with. `Resource: "*"`
Cette politique suppose l'authentification AWS SigV4. Si le principal s'authentifie avec une clé d'API, accordez-la également `aws-external-anthropic:CallWithBearerToken` (voir Authentification).

## Exemple : isolation de l'espace de travail par client

Limite un rôle à un seul espace de travail :

```
{
  "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**

Le `aws-external-anthropic:*` caractère générique de la première instruction inclut des actions limitées au compte (`CreateWorkspace`,`ListWorkspaces`) que la contrainte ARN de l'espace de travail filtre silencieusement. Cela est conforme à l'intention d' « isolation » (le rôle ne peut pas créer ou énumérer des espaces de travail), mais la politique contient des autorisations qui n'ont aucun effet. Voir Automatisation du provisionnement pour le modèle délimité par compte.

La deuxième instruction autorise `CallWithBearerToken` et `AssumeConsole` sur toutes les ressources, car les deux sont des actions sans itinéraire qui ne sont pas liées à un ARN d'espace de travail. Omettez la deuxième déclaration si le rôle utilise uniquement Sigv4 et ne se fédère jamais avec la console Claude.

## Exemple : verrouillage des fonctionnalités d'un ZDR-sensitive espace de travail

Bloque le traitement par lots et le téléchargement de fichiers sur un espace de travail spécifique tout en laissant la possibilité d'inférence synchrone. Utile lorsqu'un espace de travail gère des données ZDR (Zero Data Retention) qui ne doivent pas être conservées côté serveur. Associez cette politique à une politique d'autorisation telle que l'exemple `AnthropicLimitedAccess` d'espace de travail unique ci-dessus ; à elle seule, une Deny-only politique n'accorde aucune autorisation :

```
{
  "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**

Ce refus bloque uniquement la création. Les autres actions sur les fichiers et les lots ne sont pas refusées, sauf si vous les listez également. Pour un verrouillage complet dans lequel l'espace de travail ne doit jamais contenir de fichiers ou de lots, refusez également `aws-external-anthropic:GetFileaws-external-anthropic:ListFiles`,`aws-external-anthropic:DeleteFile`,`aws-external-anthropic:GetBatchInference`,`aws-`

`external-anthropic:ListBatchInferences`,`aws-external-anthropic:CancelBatchInference`, et`aws-external-anthropic:DeleteBatchInference`.

## Exemple : automatisation du provisionnement

Accorde à un CI/CD rôle les actions nécessaires pour créer et gérer des espaces de travail, sans aucune autorisation d'inférence :

```
{
  "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`et `ListWorkspaces` sont des opérations limitées à un compte. La spécification d'un ARN d'espace de travail pour ces actions n'a aucun effet ; utilisez`Resource: "*"`.

## Politiques gérées

AWS fournit des politiques gérées pour les modèles d'accès courants :

- **`AnthropicFullAccess`**: Subventions `aws-external-anthropic:*` pour toutes les ressources.
- **`AnthropicReadOnlyAccess`**: Subventions `Get*List*`, et `CallWithBearerToken` sur toutes les ressources.

- **`AnthropicInferenceAccess`**: accorde les ReadOnly actions ainsi que les actions d'inférence (`CreateInference`,,`CreateBatchInference`, `CancelBatchInferenceDeleteBatchInference`,`CountTokens`) sur toutes les ressources.
- **`AnthropicLimitedAccess`**: accorde les `AnthropicInferenceAccess` actions ainsi que toutes les actions des agents gérés par Claude (agents, sessions, environnements, coffres-forts, mémoires) sur toutes les ressources.
- **`AnthropicSelfHostedEnvironmentAccess`**: accorde les autorisations dont un utilisateur du sandbox auto-hébergé a besoin pour interroger et traiter les éléments de travail liés à l'environnement, lire les ressources associées à l'environnement, à la session et aux compétences, et mettre à jour l'état de la session. Associez cette politique au rôle IAM assumé par votre agent de l'environnement auto-hébergé. `AnthropicSelfHostedEnvironmentAccess`est le rôle unique par défaut recommandé ; si vous exécutez l'analyseur de travail et le sandbox par session selon des principes IAM distincts, vous pouvez répartir ces autorisations entre deux politiques personnalisées plus étroites pour le moindre privilège.

`AnthropicInferenceAccess`est la politique gérée la plus étroite suffisante pour exécuter une inférence. Grâce aux `List*` caractères génériques `Get*` et, il accorde un accès en lecture à toutes les ressources d'API de l'espace de noms, y compris le téléchargement du contenu des fichiers `GetFile` et le contenu de la mémoire. `GetMemoryStore` Il n'autorise pas la création ou la suppression de fichiers ou de compétences, la gestion des profils utilisateur, la mutation de l'espace de travail ou la fédération de consoles.

> **Note**
>
> `AnthropicReadOnlyAccessAnthropicInferenceAccess`, et `AnthropicLimitedAccess` n'accordez pas`AssumeConsole`. Les directeurs qui doivent se fédérer vers la console Claude ont besoin d'une subvention distincte, soit par le biais d'une politique personnalisée, `aws-external-anthropic:AssumeConsole` `AnthropicFullAccess` soit par le biais d'une politique personnalisée. Reportez-vous à la section Fédération avec la console Claude.

> **Note**
>
> `CreateInferenceet CreateBatchInference` sont des actions distinctes. Refuser l'un ne bloque pas l'autre. Si vous avez l'intention d'empêcher tous les appels de modèles, refusez les deux.

## Fédérer avec la console Claude

`aws-external-anthropic:AssumeConsolepermet` à un directeur IAM de se fédérer dans la console Anthropic-operated Claude. L'accès au sein de la console est toujours régi par IAM pour la plupart des opérations, mais un sous-ensemble d'opérations d'administration (principalement les vues d'utilisation qui n'ont aucune API IAM correspondante) est limité à la fonctionnalité avec laquelle le principal est fédéré.

Il existe deux fonctionnalités :

- **`developer`**— permet les opérations dont un développeur de la plateforme Claude sur AWS a besoin dans le cadre de son travail quotidien : exécution d'inférences depuis la console, lecture des données de l'espace de travail, visualisation de l'utilisation personnelle.
- **`admin`**— autorise en outre les opérations de console réservées aux administrateurs, y compris les vues d'utilisation à l'échelle du compte et les paramètres administratifs qui ne sont pas visibles lors des actions IAM.

Contrôlez la capacité qu'un principal peut demander à l'aide de la clé de `aws-external-anthropic:Capability` condition de l'`AssumeConsole`action :

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

```
    }
  ]
}
```

`AnthropicFullAccess`subventions `AssumeConsole` sans restriction de capacité. Pour toute autorisation plus limitée (accès à la console réservé aux développeurs ou accès réservé aux administrateurs), associez une politique personnalisée qui s'applique à la condition indiquée ci-dessus. `AssumeConsole` `aws-external-anthropic:Capability`

## Téléphoner avec des jetons au porteur

`aws-external-anthropic:CallWithBearerToken`autorise le chemin de SigV4-free requête utilisé lorsqu'une clé d'API est présentée sous forme de jeton porteur. Tout principal qui s'authentifie à l'aide d'une clé d'API a besoin de cette action sur l'espace de travail cible. Cela s'applique que vous utilisiez `ANTHROPIC_AWS_API_KEY` ou définissiez directement l'`Authorization: Bearer`en-tête.

Cela est requis pour les appelants aux clés d'API en plus des actions d'inférence (`CreateInference`,`CreateBatchInference`, etc.). Sans`CallWithBearerToken`, les demandes de clé d'API sont rejetées avant d'atteindre le contrôle d'autorisation d'inférence. Les appelants SIGv4 n'ont pas besoin de cette action.

`AnthropicReadOnlyAccess`,`AnthropicInferenceAccess`,`AnthropicLimitedAccess`, et `AnthropicFullAccess` tout inclut`CallWithBearerToken`. Si vous rédigez une politique personnalisée pour l'accès aux clés d'API, ajoutez-la explicitement.

## Fédération d'identité Web sortante (requise pour l'accès à la console)

La console Claude fonctionne dans l'infrastructure Anthropic, et non dans AWS. Lorsqu'un responsable IAM appelle`AssumeConsole`, AWS STS émet un jeton d'identité Web destiné au public d'Anthropic ; la console Anthropic accepte ensuite ce jeton et établit la session fédérée. Pour que cela fonctionne, le compte AWS doit autoriser la fédération d'identité Web sortante au `aws-external-anthropic` public.

Les principaux qui appellent `AssumeConsole` ont besoin des autorisations STS suivantes en plus de l'`aws-external-anthropic:AssumeConsole`action :

- **`sts:GetWebIdentityToken`**— permet d'émettre le jeton d'identité Web consommé par la console Anthropic.
- **`sts:TagGetWebIdentityToken`**— permet de joindre des balises de session au jeton d'identité Web. Claude Platform sur AWS utilise ces balises pour transmettre les capacités du directeur et le contexte de l'espace de travail à la console.

Incluez les deux dans la politique de confiance de tout rôle assumé par les utilisateurs de la console ou dans la inline/managed politique attachée aux identités des utilisateurs qui appellent `AssumeConsole` directement. `AnthropicFullAccess`inclut les deux actions STS. Les environnements qui refusent `sts:*` au niveau du SCP ou de la limite d'autorisation doivent autoriser explicitement ces deux actions pour que la fédération de consoles réussisse.

# Actions IAM pour Claude Platform sur AWS

Référence d'action IAM pour contrôler l'accès à la plateforme Claude sur AWS via les politiques AWS.

Claude Platform sur AWS utilise AWS IAM pour le contrôle d'accès. Chaque itinéraire d'API correspond à une action IAM dans l'espace de `aws-external-anthropic` noms. Cette page répertorie toutes les actions, les itinéraires autorisés par chaque action et les politiques gérées disponibles pour les modèles d'accès courants. Pour la configuration et l'authentification de la plateforme, consultez Qu'est-ce que la plateforme Claude sur AWS ? .

## Description détaillée du service

| | |
|---|---|
| Préfixe de service IAM | `aws-external-anthropic` |
| Resource types | `workspace` |
| ARN de l'espace de travail | `arn:aws:aws-external-anthropic:{region}:{account-id}:workspace/{workspace-id}` |

La région ARN est toujours renseignée et correspond à la région à laquelle l'espace de travail est lié. Le segment de ressource est l'ID d'espace de travail étiqueté (`wrkspc_…` ), la même valeur que celle que vous transmettez dans l'`anthropic-workspace-id`en-tête.

## Actions

Le service définit 65 actions réparties dans 15 groupes. Les actions respectent la `VerbNoun` convention AWS et utilisent la discipline verbale afin que `Get*` les `List*` caractères génériques produisent une limite claire en lecture seule.

### Inférence

| Action | Routes autorisées |
|---|---|
| `CreateInference` | `POST /v1/messages` |

| Action | Routes autorisées |
| --- | --- |
| CountTokens | POST /v1/messages/count_tokens |

## Traitement par lots

| Action | Routes autorisées |
| --- | --- |
| 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} |

## Modèles

| Action | Routes autorisées |
| --- | --- |
| GetModel | GET /v1/models/{id} |
| ListModels | GET /v1/models |

## Fichiers

| Action | Routes autorisées |
| --- | --- |
| CreateFile | POST /v1/files |
| GetFile | GET /v1/files/{id} GET /v1/files/{id}/content |
| ListFiles | GET /v1/files |

| Action | Routes autorisées |
| --- | --- |
| `DeleteFile` | `DELETE /v1/files/{id}` |

> **Note**
>
> `GetFile`autorise à la fois le téléchargement des métadonnées et du contenu. Un principal disposant d'un accès en lecture seule peut télécharger des octets de fichiers, et pas simplement des fichiers de liste.

## Compétences

| Action | Routes autorisées |
| --- | --- |
| `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**
>
> La suppression d'une version de compétence individuelle correspond à`UpdateSkill`, non`DeleteSkill`. Une politique qui refuse autorise `aws-external-anthropic:Delete*` toujours la suppression de version. `UpdateSkill`Refusez et `CreateSkill` aussi si vous devez empêcher toute mutation de compétence.

## Profils utilisateurs

| Action | Routes autorisées |
| --- | --- |
| `CreateUserProfile` | `POST /v1/user_profiles` |
| `GetUserProfile` | `GET /v1/user_profiles/{id}` |
| `ListUserProfiles` | `GET /v1/user_profiles` |
| `UpdateUserProfile` | `POST /v1/user_profiles/{id}` |

> **Warning**
>
> La correspondance des actions IAM ne fait pas la distinction majuscules/majuscules. Le caractère générique `aws-external-anthropic:*File` correspond à `CreateFile`Get`File`, et`DeleteFile`, mais ne correspond pas `ListFiles` (ce qui se termine par « fichiers » et non par « fichier »). Il surcorrespond également `CreateUserProfile`Get`UserProfile`, et `UpdateUserProfile` parce que « Profile » se termine par « file ». Si vous avez l'intention d'autoriser ou de refuser uniquement les actions de l'API Files, énumérez-les explicitement (`CreateFile`,, `GetFile`List`Files`,`DeleteFile`) plutôt que d'utiliser un modèle de `*File` suffixe.

## Agents

Définitions des agents pour Claude Managed Agents.

| Action | Routes autorisées |
| --- | --- |
| `CreateAgent` | L'agent crée des itinéraires |
| `GetAgent` | L'agent lit les itinéraires |
| `ListAgents` | Liste des itinéraires des agents |
| `UpdateAgent` | Itinéraires de mise à jour de |

| Action | Routes autorisées |
| --- | --- |
| `ArchiveAgent` | Routes d'archivage des agents |

## Séances

Historique des sessions et des événements avec les agents.

| Action | Routes autorisées |
| --- | --- |
| `CreateSession` | Session : créer des itinéraires |
| `GetSession` | Itinéraires de lecture des sessions |
| `ListSessions` | Itinéraires de la liste des sessions |
| `UpdateSession` | itinéraires de mise à jour de session |
| `ArchiveSession` | Routes d'archivage des sessions |
| `DeleteSession` | Session : supprimer des itinéraires |

## Environnements

Configurations de conteneurs cloud pour les sessions.

| Action | Routes autorisées |
| --- | --- |
| `CreateEnvironment` | Environnement : créer des itinéraires |
| `GetEnvironment` | Environnement : lisez les itinéraires |
| `ListEnvironments` | Itinéraires de la liste d'environnement |
| `UpdateEnvironment` | Itinéraires de mise à jour environnementale |
| `ArchiveEnvironment` | Routes d'archivage de l'environnement |

| Action | Routes autorisées |
| --- | --- |
| `DeleteEnvironment` | Environnement : supprimer des itinéraires |
| `ProcessEnvironmentWork` | Self-hosted itinéraires des travailleurs de l'environnement |

## Coffres-forts

Coffres-forts d'informations d'identification pour l'authentification de session.

| Action | Routes autorisées |
| --- | --- |
| `CreateVault` | Vault crée des itinéraires |
| `GetVault` | Vault lit les itinéraires |
| `ListVaults` | Liste des itinéraires du coffre-fort |
| `UpdateVault` | itinéraires de mise à jour du coffre-fort |
| `ArchiveVault` | Routes d'archivage du coffre-fort |
| `DeleteVault` | Supprimer des itinéraires dans le coffre-fort |

> **Note**
>
> Les opérations du coffre-fort sont classées comme CloudTrail des événements de gestion (plutôt que comme des événements de données) car les coffres-forts contiennent des informations d'identification et bénéficient de la journalisation des audits par défaut. Les autres opérations des agents gérés par Claude (agents, sessions, environnements, mémoires) sont des événements de données.

## Magasins de mémoire

Mémoire persistante de l'agent.

| Action | Routes autorisées |
| --- | --- |
| `CreateMemoryStore` | Memory Store crée des itinéraires |
| `GetMemoryStore` | Mémoriser les itinéraires de lecture |
| `ListMemoryStores` | Itinéraires de liste de stockage de mémoire |
| `UpdateMemoryStore` | Itinéraires de mise à jour de la mémoire |
| `ArchiveMemoryStore` | Routes d'archivage de la mémoire |
| `DeleteMemoryStore` | Stockage de mémoire : supprimer des itinéraires |

> **Note**
>
> `GetMemoryStore`lit le contenu de la mémoire. Le `Get*` caractère générique d'une politique gérée ou personnalisée accorde donc un accès en lecture à la mémoire. Appliquez les politiques de manière explicite si le contenu de la mémoire doit être restreint.

## Webhooks

Notifications d'événements du cycle de vie pour les sessions.

| Action | Routes autorisées |
| --- | --- |
| `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 | Routes autorisées |
| --- | --- |
| `RotateWebhookSecret` | `POST /v1/webhooks/{webhook_endpoint_id}/regenerate_signing_secret` |

## Espaces de travail

| Action | Routes autorisées |
| --- | --- |
| `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` |

Un espace de travail par défaut est fourni lors de l'inscription ; voir Espaces de travail.

## Authentification

| Action | Routes autorisées |
| --- | --- |
| `CallWithBearerToken` | (aucun) |

`CallWithBearerToken`est une autorisation de couche d'authentification qui autorise un mandant à s'authentifier via une clé d'API (jeton porteur) plutôt qu'avec AWS Sigv4. Il ne correspond pas à un itinéraire. Accordez-le en même temps que les actions cartographiées que vous souhaitez que le détenteur de la clé d'API effectue.

# Accès à la console

| Action | Routes autorisées |
| --- | --- |
| `AssumeConsole` | (aucun) |

`AssumeConsole`autorise un directeur à ouvrir la console Claude pour une plateforme Claude sur un espace de travail AWS via le flux de fédération de la console AWS. Il ne correspond pas à un itinéraire. Accordez-le aux directeurs qui devraient pouvoir cliquer sur Ouvrir la console Claude sur la page du service Claude Platform on AWS dans la console AWS. La clé de `aws-external-anthropic:Capability` condition de l'`AssumeConsole`action contrôle la capacité de console (développeur ou administrateur) qu'un mandant peut demander. Consultez les politiques IAM pour plus de détails et l'utilisation de la console Claude pour le flux de connexion.

**Warning**

`aws-external-anthropic:AssumeConsole`émet une session Claude Console d'une durée maximale de 12 heures, indépendamment de la durée de la session de l'appelant. Un appelant dont la session IAM dure moins de 12 heures peut toujours obtenir une session de console dont la durée est supérieure à celle de ses informations d'identification source. Limitez cette autorisation aux principaux qui ont besoin d'accéder à la console Claude, et révoquez-la rapidement lorsqu'elle n'est plus nécessaire.

**Note**

Les actions effectuées dans la console Claude après la fédération ne sont pas enregistrées dans AWS CloudTrail et ne sont pas attribuables via IAM. Cela inclut les activités globales liées à l'espace de travail, telles que la consultation des rapports d'utilisation. Si vous avez besoin d'une piste d'audit pour l'activité de la console, utilisez plutôt les journaux d'audit disponibles dans la console Claude CloudTrail.

# Route-to-action cartographie

Le tableau suivant répertorie tous les itinéraires sur la plateforme Claude sur AWS et l'action IAM requise pour l'appeler. L'action IAM de la route stable autorise également les demandes qui utilisent l'`anthropic-beta`en-tête. CloudTrail classe chaque action en tant qu'événement de données (volumes élevés, opérations sur le plan de données) ou en événement de gestion (opérations sur le plan de contrôle).

| Méthode | Acheminement | Action IAM | CloudTrail type d'événement |
|---|---|---|---|
| `POST` | `/v1/messages` | `CreateInference` | Données |
| `POST` | `/v1/messages/count_tokens` | `CountTokens` | Données |
| `POST` | `/v1/messages/batches` | `CreateBatchInference` | Données |
| `GET` | `/v1/messages/batches` | `ListBatchInferences` | Données |
| `GET` | `/v1/messages/batches/{id}` | `GetBatchInference` | Données |
| `GET` | `/v1/messages/batches/{id}/results` | `GetBatchInference` | Données |
| `POST` | `/v1/messages/batches/{id}/cancel` | `CancelBatchInference` | Données |
| `DELETE` | `/v1/messages/batches/{id}` | `DeleteBatchInference` | Données |
| `GET` | `/v1/models` | `ListModels` | Données |
| `GET` | `/v1/models/{id}` | `GetModel` | Données |
| `POST` | `/v1/files` | `CreateFile` | Données |

| Méthode | Acheminement | Action IAM | CloudTrail type d'événement |
|---|---|---|---|
| GET | /v1/files | ListFiles | Données |
| GET | /v1/files/{id} | GetFile | Données |
| GET | /v1/files/{id}/content | GetFile | Données |
| DELETE | /v1/files/{id} | DeleteFile | Données |
| POST | /v1/skills | CreateSkill | Données |
| GET | /v1/skills | ListSkills | Données |
| GET | /v1/skills/{id} | GetSkill | Données |
| DELETE | /v1/skills/{id} | DeleteSkill | Données |
| POST | /v1/skills/{id}/versions | UpdateSkill | Données |
| GET | /v1/skills/{id}/versions | GetSkill | Données |
| GET | /v1/skills/{id}/versions/{version} | GetSkill | Données |
| GET | /v1/skills/{id}/versions/{version}/content | GetSkill | Données |
| DELETE | /v1/skills/{id}/versions/{version} | UpdateSkill | Données |
| POST | /v1/user_profiles | CreateUserProfile | Données |
| GET | /v1/user_profiles | ListUserProfiles | Données |
| GET | /v1/user_profiles/{id} | GetUserProfile | Données |
| POST | /v1/user_profiles/{id} | UpdateUserProfile | Données |

| Méthode | Acheminement | Action IAM | CloudTrail type d'événement |
|---|---|---|---|
| POST | /v1/organizations/workspaces | CreateWorkspace | Gestion |
| GET | /v1/organizations/workspaces | ListWorkspaces | Gestion |
| GET | /v1/organizations/workspaces/{id} | GetWorkspace | Gestion |
| POST | /v1/organizations/workspaces/{id} | UpdateWorkspace | Gestion |
| POST | /v1/organizations/workspaces/{id}/archive | ArchiveWorkspace | Gestion |

Les itinéraires des agents gérés par Claude (agents, sessions, environnements, coffres-forts, mémoires) suivent le même schéma de cheminement vers l'action ; pour le mappage complet par itinéraire, consultez la référence des actions Anthropic IAM. Les itinéraires du coffre-fort sont des événements de gestion ; les itinéraires d'agent, de session, d'environnement et de stockage de mémoire sont des événements de données.

Les itinéraires ne figurant pas sur la plateforme Claude sur AWS sont refusés par défaut à la passerelle.

## Consultez aussi

- Les politiques IAM, par exemple les politiques et les politiques gérées
- Guide de l'utilisateur AWS IAM pour la syntaxe des politiques IAM et la logique d'évaluation
- Guide de CloudTrail l'utilisateur AWS pour la configuration de la journalisation des audits

# Historique du document

Le tableau suivant décrit les modifications importantes apportées au guide de l'utilisateur de la plateforme Claude sur AWS.

| Modifier | Description | Date |
| --- | --- | --- |
| Publication initiale | Publication du guide de l'utilisateur de la plateforme Claude sur AWS. | 7 mai 2026 |

Les traductions sont fournies par des outils de traduction automatique. En cas de conflit entre le contenu d'une traduction et celui de la version originale en anglais, la version anglaise prévaudra.