AWS Livre blanc

# SageMaker Bonnes pratiques en matière d'administration du studio

Copyright © 2024 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

# SageMaker Bonnes pratiques en matière d'administration du studio: AWS Livre blanc

Copyright © 2024 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Les marques et la présentation commerciale d’Amazon ne peuvent être utilisées en relation avec un produit ou un service qui n’est pas d’Amazon, d’une manière susceptible de créer une confusion parmi les clients, ou d’une manière qui dénigre ou discrédite Amazon. Toutes les autres marques commerciales qui ne sont pas la propriété d’Amazon appartiennent à leurs propriétaires respectifs, qui peuvent ou non être affiliés ou connectés à Amazon, ou sponsorisés par Amazon.

# Table of Contents

Résumé et introduction ..... i
Résumé ..... 1
Êtes-vous Well-Architected ? ..... 1
Introduction ..... 1
Modèle d'exploitation ..... 3
Structure de compte recommandée ..... 3
Modèle de structure de compte centralisé ..... 4
Modèle de structure de compte décentralisé ..... 5
Modèle de structure de compte fédéré ..... 6
Plateforme ML, mutualisation ..... 7
Gestion de domaine ..... 9
Domaines multiples et espaces partagés ..... 11
Configurez des espaces partagés dans votre domaine ..... 12
Configurez votre domaine IAM (pour) la fédération ..... 12
Configurez votre domaine pour la fédération d'authentification unique (SSO) ..... 12
SageMaker Profil utilisateur d'AI Studio ..... 13
Application Jupyter Server ..... 13
L'application Jupyter Kernel Gateway ..... 13
EFSVolume d'Amazon ..... 14
Sauvegarde et restauration ..... 15
EBSVolume d'Amazon ..... 15
Sécurisation de l'accès aux documents pré-signés URL ..... 16
SageMaker Quotas et limites de domaines AI ..... 17
Gestion des identités ..... 19
Utilisateurs, groupes et rôles ..... 19
Fédération d'utilisateurs ..... 21
Utilisateurs IAM ..... 21
AWS IAMou fédération de comptes ..... 22
SAMLauthentification à l'aide de AWS Lambda ..... 23
AWSIAMFédération iDC ..... 24
Conseils d'authentification de domaine ..... 25
Gestion des autorisations ..... 26
Rôles et stratégies IAM ..... 26
SageMaker Flux de travail d'autorisation d'AI Studio Notebook ..... 28

IAMFédération : flux de travail de Studio Notebook ..... 28
Environnement déployé : flux de formation SageMaker basé sur l'IA ..... 29
Autorisations relatives aux données ..... 30
Accès aux AWS Lake Formation données ..... 30
Rambardes communes ..... 32
Limitez l'accès au bloc-notes à des instances spécifiques ..... 33
Limiter les domaines SageMaker AI Studio non conformes ..... 33
Limitez le lancement d'images SageMaker IA non autorisées ..... 34
Lancez des blocs-notes uniquement via des points de terminaison basés sur SageMaker l'IA VPC ..... 35
Limitez SageMaker l'accès aux ordinateurs portables AI Studio à une plage d'adresses IP limitée ..... 35
Empêcher les utilisateurs d' SageMaker AI Studio d'accéder à d'autres profils utilisateur ..... 36
Appliquer le balisage ..... 37
Accès root dans SageMaker AI Studio ..... 38
Gestion du réseau ..... 40
VPCplanification du réseau ..... 40
VPCoptions réseau ..... 42
Limites ..... 44
Protection des données ..... 45
Protégez les données au repos ..... 45
Chiffrement au repos avec AWS KMS ..... 46
Protéger les données en transit ..... 46
Garde-corps de protection des données ..... 47
Chiffrez les volumes d'hébergement SageMaker AI au repos ..... 47
Chiffrer les compartiments S3 utilisés lors de la surveillance des modèles ..... 47
Chiffrer un volume de stockage de domaine SageMaker AI Studio ..... 48
Chiffrez les données stockées dans S3 qui sont utilisées pour partager des blocs-notes ..... 49
Limites ..... 49
Journalisation et surveillance ..... 50
Se connecter avec CloudWatch ..... 50
Audit avec AWS CloudTrail ..... 53
Attribution des coûts ..... 55
Marquage automatique ..... 55
Suivi des coûts ..... 55
Contrôle des coûts ..... 56

Personnalisation ..... 58
Configuration du cycle de vie ..... 58
Images personnalisées pour les ordinateurs portables SageMaker AI Studio ..... 58
JupyterLab extensions ..... 59
Référentiels Git ..... 59
Environnement Conda ..... 60
Conclusion ..... 61
Annexe ..... 62
Comparaison entre plusieurs locataires ..... 62
SageMaker Sauvegarde et restauration de domaines AI Studio ..... 63
Option 1 : Sauvegarder à partir d'une sauvegarde existante à EFS l'aide d'un EC2 ..... 64
Option 2 : sauvegarde à partir de données existantes à EFS l'aide de S3 et de la configuration du cycle de vie ..... 65
SageMaker Accès au studio à l'aide d'une SAML assertion ..... 66
Suggestions de lecture ..... 68
Collaborateurs ..... 69
Révisions du document ..... 70
Avis ..... 71
Glossaire AWS ..... 72
..... lxxiii

# SageMaker Bonnes pratiques en matière d'administration du studio

Date de publication : 25 avril 2023 (Révisions du document)

## Résumé

Amazon SageMaker AI Studio fournit une interface visuelle Web unique dans laquelle vous pouvez effectuer toutes les étapes de développement du machine learning (ML), ce qui améliore la productivité des équipes de data science. SageMaker AI Studio vous offre un accès, un contrôle et une visibilité complets sur chaque étape requise pour créer, former et évaluer des modèles.

Dans ce livre blanc, nous discutons des meilleures pratiques sur des sujets tels que le modèle d'exploitation, la gestion des domaines, la gestion des identités, la gestion des autorisations, la gestion du réseau, la journalisation, la surveillance et la personnalisation. Les meilleures pratiques décrites ici sont destinées au déploiement d' SageMaker AI Studio en entreprise, y compris les déploiements multi-locataires. Ce document est destiné aux administrateurs de plateformes ML, aux ingénieurs ML et aux architectes ML.

## Êtes-vous Well-Architected ?

Le AWS Well-Architected Framework vous aide à comprendre les avantages et les inconvénients des décisions que vous prenez lors de la création de systèmes dans le cloud. Les six piliers du cadre vous permettent d'apprendre les meilleures pratiques architecturales pour concevoir et exploiter des systèmes fiables, sécurisés, efficaces, rentables et durables. À l'aide du AWS Well-Architected Tool, disponible gratuitement dans le AWS Management Console, vous pouvez évaluer votre charge de travail par rapport à ces meilleures pratiques en répondant à une série de questions pour chaque pilier.

Dans le Machine Learning Lens, nous nous concentrons sur la manière de concevoir, déployer et structurer vos charges de travail d'apprentissage automatique dans le AWS Cloud. Cet objectif s'ajoute aux meilleures pratiques décrites dans le Well-Architected Framework.

## Introduction

Lorsque vous administrez SageMaker AI Studio en tant que plateforme de ML, vous avez besoin de conseils sur les meilleures pratiques pour prendre des décisions éclairées afin de vous aider

à adapter votre plateforme de ML à mesure que vos charges de travail augmentent. Pour le provisionnement, l'opérationnalisation et le dimensionnement de votre plateforme ML, tenez compte des points suivants :

- Choisissez le bon modèle d'exploitation et organisez vos environnements de machine learning pour atteindre vos objectifs commerciaux.
- Choisissez comment configurer l'authentification de domaine SageMaker AI Studio pour les identités des utilisateurs et tenez compte des limites au niveau du domaine.
- Décidez comment fédérer l'identité et l'autorisation de vos utilisateurs à la plateforme ML pour des contrôles d'accès et des audits précis.
- Envisagez de configurer des autorisations et des garde-fous pour les différents rôles de vos personas ML.
- Planifiez la topologie de votre réseau de cloud privé virtuel (VPC) en tenant compte de la sensibilité de votre charge de travail ML, du nombre d'utilisateurs, des types d'instances, des applications et des tâches lancées.
- Classez et protégez vos données au repos et en transit grâce au chiffrement.
- Réfléchissez à la manière de consigner et de surveiller les différentes interfaces de programmation d'applications (APIs) et les activités des utilisateurs à des fins de conformité.
- Personnalisez l'expérience d' SageMaker AI Studio Notebook avec vos propres images et scripts de configuration du cycle de vie.

# Modèle d'exploitation

Un modèle opérationnel est un cadre qui réunit les personnes, les processus et les technologies pour aider une organisation à générer de la valeur commerciale de manière évolutive, cohérente et efficace. Le modèle opérationnel ML fournit un processus de développement de produits standard pour les équipes de l'organisation. Il existe trois modèles de mise en œuvre du modèle opérationnel, en fonction de la taille, de la complexité et des facteurs commerciaux :

- Équipe de science des données centralisée — Dans ce modèle, toutes les activités de science des données sont centralisées au sein d'une seule équipe ou organisation. Ce modèle est similaire au modèle Center of Excellence (COE), dans lequel toutes les unités commerciales font appel à cette équipe pour des projets de science des données.
- Équipes de science des données décentralisées — Dans ce modèle, les activités de science des données sont réparties entre différentes fonctions ou divisions commerciales, ou basées sur différentes gammes de produits.
- Équipes de data science fédérées — Dans ce modèle, les fonctions de services partagés telles que les référentiels de code, les pipelines d'intégration continue et de livraison continue (CI/CD), etc. sont gérées par l'équipe centralisée, et chaque unité commerciale ou fonction au niveau du produit est gérée par des équipes décentralisées. Cela est similaire au modèle hub and spoke, dans lequel chaque unité commerciale dispose de ses propres équipes de science des données ; toutefois, ces équipes coordonnent leurs activités avec l'équipe centralisée.

Avant de décider de lancer votre premier domaine de studio pour des cas d'utilisation en production, réfléchissez à votre modèle d'exploitation et aux AWS meilleures pratiques en matière d'organisation de votre environnement. Pour plus d'informations, reportez-vous à la section Organisation de votre AWS environnement à l'aide de plusieurs comptes.

La section suivante fournit des conseils sur l'organisation de votre structure de compte pour chacun des modèles opérationnels.

## Structure de compte recommandée

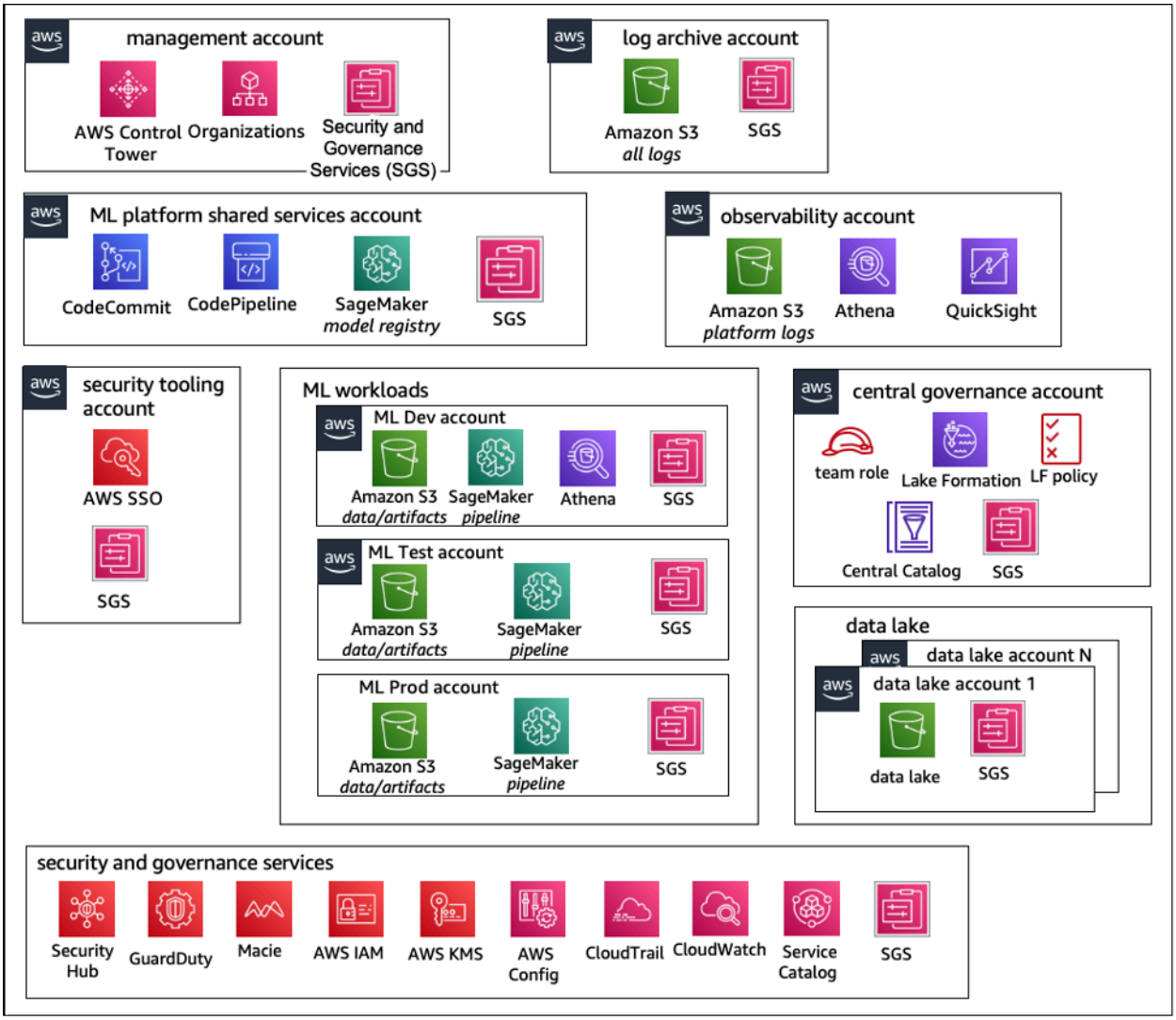
Dans cette section, nous présentons brièvement un modèle de structure de compte opérationnel que vous pouvez utiliser au départ et modifier en fonction des exigences opérationnelles de votre organisation. Quel que soit le modèle d'exploitation que vous choisissez, nous vous recommandons de mettre en œuvre les meilleures pratiques courantes suivantes :

- AWS Control TowerÀ utiliser pour la configuration, la gestion et la gouvernance de vos comptes.
- Centralisez vos identités auprès de votre fournisseur d'identité (IdP) AWS IAMet d'Identity Center avec un compte Security Tooling à administrateur délégué et sécurisez l'accès aux charges de travail.
- Exécutez les charges de travail ML en isolant les charges de travail de développement, de test et de production au niveau du compte.
- Diffusez les journaux de charge de travail ML vers un compte d'archive de journaux, puis filtrez et appliquez une analyse des journaux dans un compte d'observabilité.
- Gérez un compte de gouvernance centralisé pour le provisionnement, le contrôle et l'audit de l'accès aux données.
- Intégrez des services de sécurité et de gouvernance (SGS) dotés de dispositifs de prévention et de détection appropriés dans chaque compte afin de garantir la sécurité et la conformité, conformément aux exigences de votre organisation et de votre charge de travail.

## Modèle de structure de compte centralisé

Dans ce modèle, l'équipe de la plateforme ML est chargée de fournir :

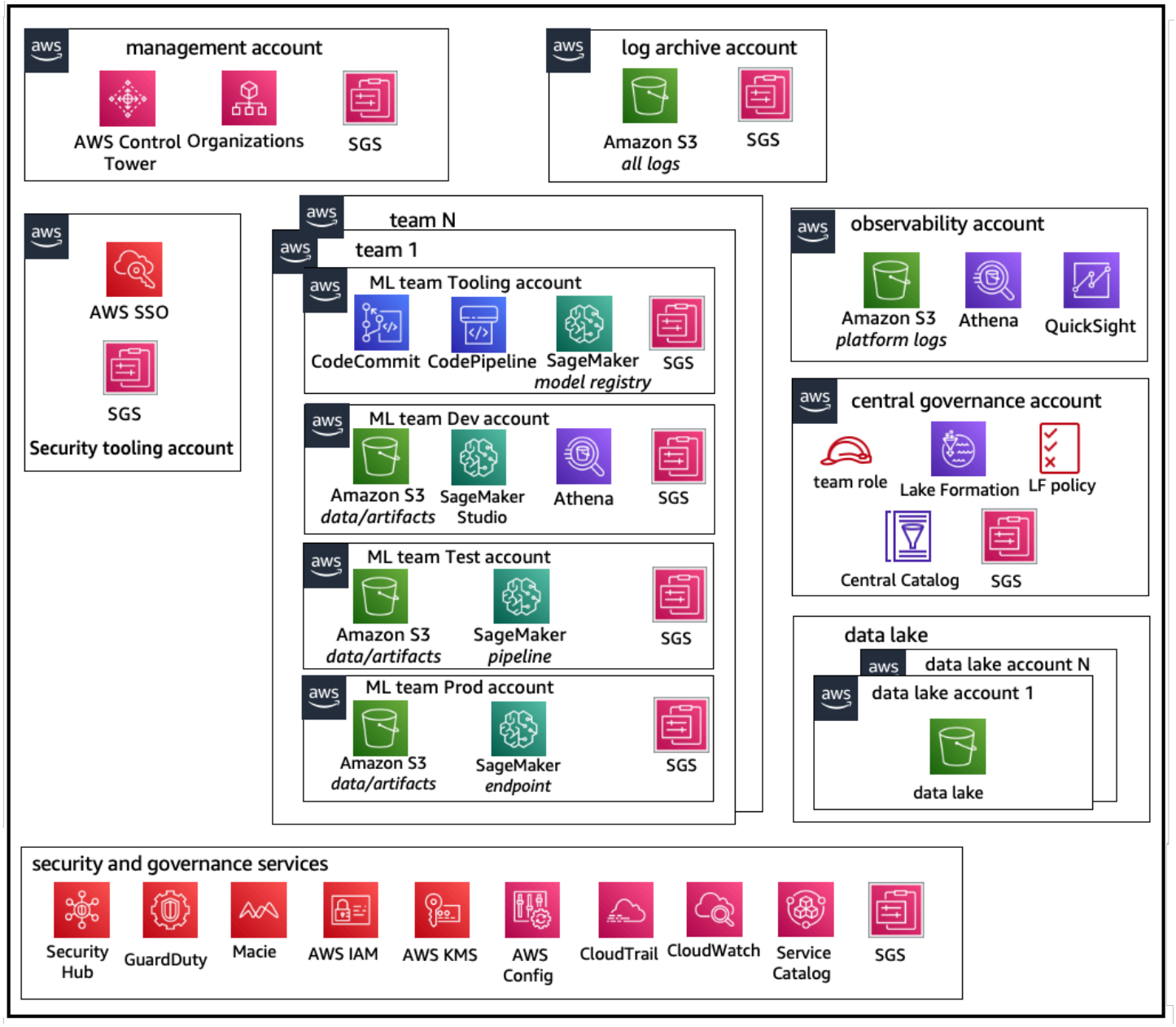
- Un compte d'outillage de services partagés qui répond aux exigences des équipes de science des données en matière d'opérations de Machine Learning (MLOps).
- Des comptes de développement, de test et de production de charges de travail ML partagés entre les équipes de data science.
- Des politiques de gouvernance garantissant que la charge de travail de chaque équipe de data science fonctionne de manière isolée.
- Bonnes pratiques courantes.




Structure de compte du modèle d'exploitation centralisé

## Modèle de structure de compte décentralisé

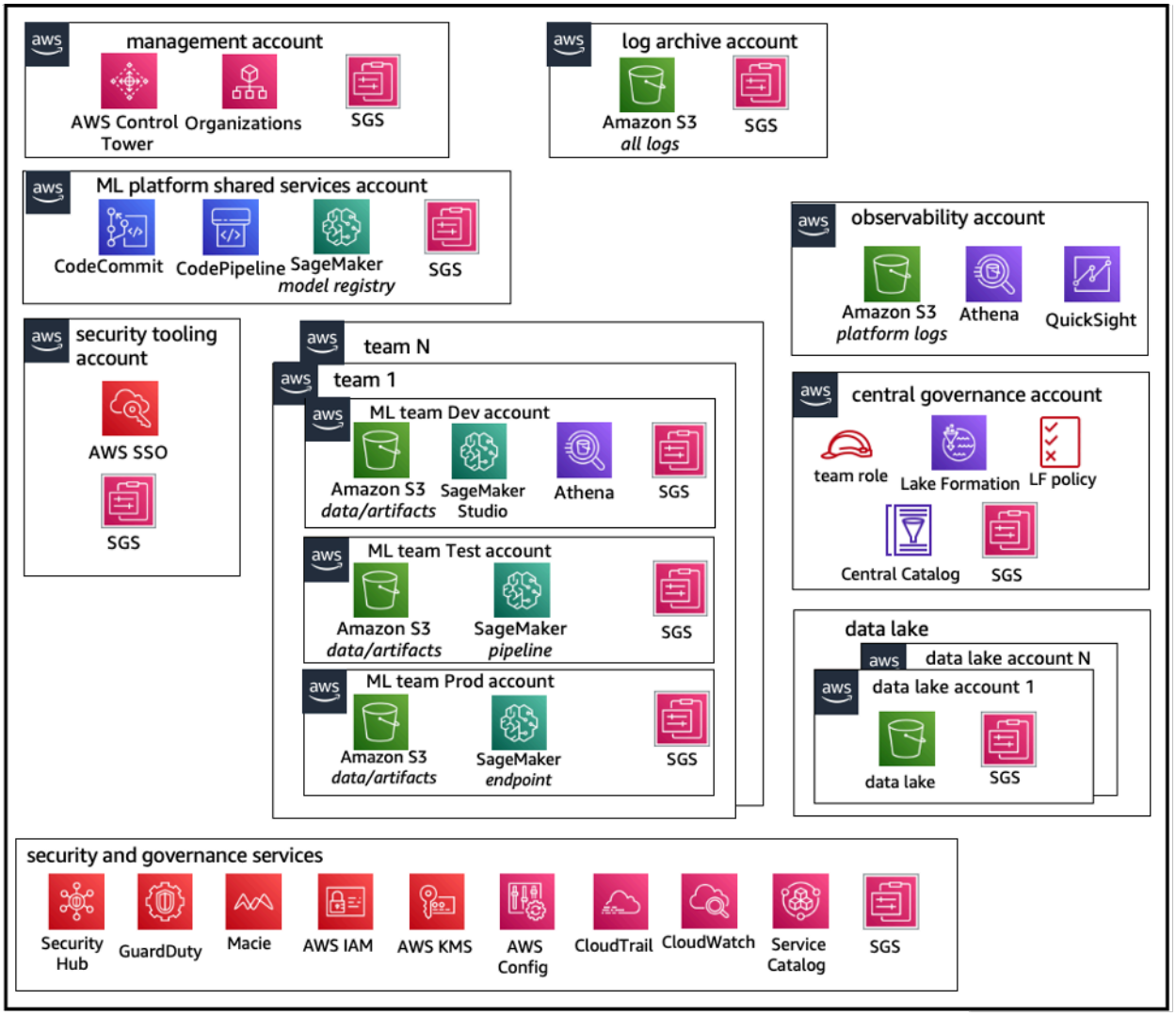
Dans ce modèle, chaque équipe de ML fonctionne de manière indépendante pour approvisionner, gérer et gouverner les comptes et les ressources de ML. Cependant, nous recommandons aux équipes de machine learning d'utiliser une approche centralisée d'observabilité et de gouvernance des données afin de simplifier la gouvernance des données et la gestion des audits.




Structure de compte du modèle opérationnel décentralisé

## Modèle de structure de compte fédéré

Ce modèle est similaire au modèle centralisé ; toutefois, la principale différence réside dans le fait que chaque science/ML team gets their own set of development/test/production charge de travail de données permet une isolation physique robuste de ses ressources de machine learning et permet également à chaque équipe d'évoluer indépendamment sans impact sur les autres équipes.




Structure de compte du modèle d'exploitation fédéré

# Plateforme ML, mutualisation

La mutualisation est une architecture logicielle dans laquelle une seule instance logicielle peut desservir plusieurs groupes d'utilisateurs distincts. Un locataire est un groupe d'utilisateurs qui partagent un accès commun avec des privilèges spécifiques à l'instance logicielle. Par exemple, si vous créez plusieurs produits ML, chaque équipe produit ayant des exigences d'accès similaires peut être considérée comme un locataire ou une équipe.

Bien qu'il soit possible de mettre en œuvre plusieurs équipes au sein d'une instance SageMaker AI Studio (telle que SageMaker AI Domain), évaluez ces avantages par rapport à des compromis tels que le rayon d'explosion, l'attribution des coûts et les limites de niveau de compte lorsque vous

regroupez plusieurs équipes dans un seul domaine SageMaker AI Studio. Pour en savoir plus sur ces compromis et les meilleures pratiques, consultez les sections suivantes.

Si vous avez besoin d'une isolation absolue des ressources, pensez à implémenter des domaines SageMaker AI Studio pour chaque locataire d'un compte différent. En fonction de vos exigences en matière d'isolation, vous pouvez implémenter plusieurs secteurs d'activité (LOBs) sous forme de domaines multiples au sein d'un même compte et d'une même région. Utilisez des espaces partagés pour une collaboration en temps quasi réel entre les membres d'une même équipe/LOB. Avec plusieurs domaines, vous continuerez à utiliser les politiques et autorisations de gestion des accès aux identités (IAM) pour garantir l'isolation des ressources.

SageMaker Les ressources d'IA créées à partir d'un domaine sont automatiquement étiquetées avec le domaine Amazon Resource Name (ARN) et le profil ou l'espace utilisateur ARN pour isoler facilement les ressources. Pour des exemples de politiques, reportez-vous à la documentation sur l'isolation des ressources du domaine. Vous pouvez y voir la référence détaillée indiquant quand utiliser une stratégie multi-comptes ou multidomaines, ainsi que les comparaisons de fonctionnalités dans la documentation, et vous pouvez consulter des exemples de scripts pour compléter les balises des domaines existants dans le référentiel. GitHub

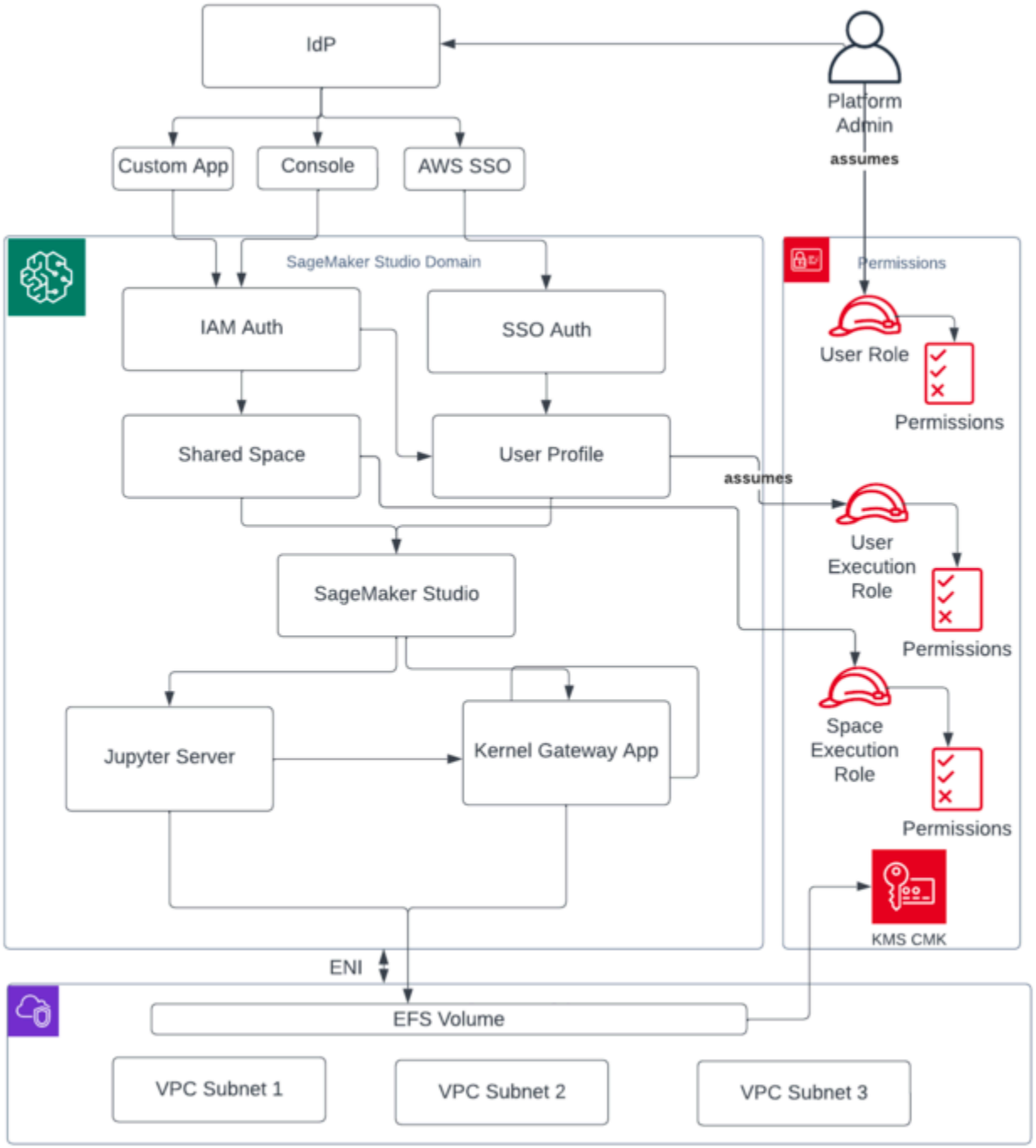
Enfin, vous pouvez implémenter un déploiement en libre-service des ressources d' SageMaker AI Studio sur plusieurs comptes à l'aide AWS Service Catalogde. Pour plus d'informations, reportez-vous à la section Gérer les AWS Service Catalog produits en plusieurs Comptes AWS et Régions AWS.

# Gestion de domaine

Un domaine Amazon SageMaker AI est composé des éléments suivants :

- Un volume Amazon Elastic File System (AmazonEFS) associé
- Liste des utilisateurs autorisés
- Une variété de configurations de sécurité, d'applications, de politiques et d'Amazon Virtual Private Cloud (AmazonVPC)

Le schéma suivant fournit une vue d'ensemble des différents composants qui constituent un SageMaker AIStudio domaine :

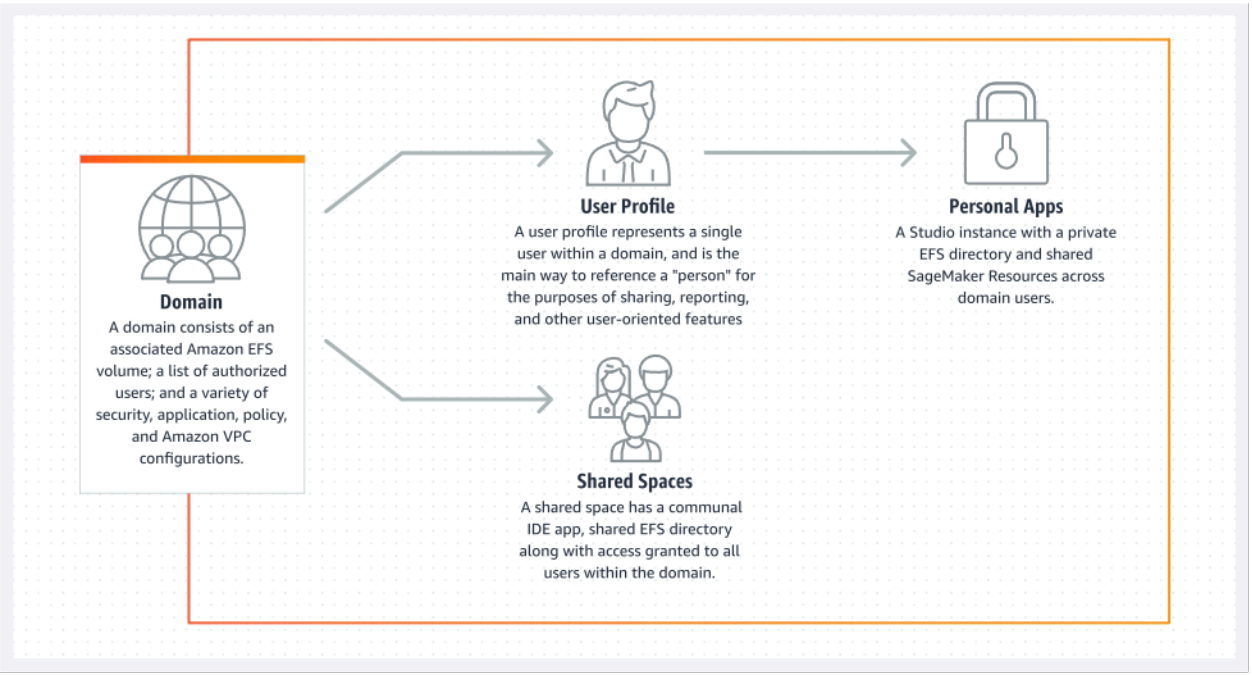



Vue de haut niveau des différents composants qui constituent un domaine SageMaker AI Studio

# Domaines multiples et espaces partagés

Amazon SageMaker AI prend désormais en charge la création de plusieurs domaines d' SageMaker IA en un seul Région AWS pour chaque compte. Chaque domaine peut avoir ses propres paramètres de domaine, tels que le mode d'authentification, et ses propres paramètres réseau, tels que VPC les sous-réseaux. Un profil utilisateur ne peut pas être partagé entre les domaines. Si un utilisateur humain fait partie de plusieurs équipes séparées par des domaines, créez un profil utilisateur pour l'utilisateur dans chaque domaine. Reportez-vous à la présentation des domaines multiples pour en savoir plus sur le remblayage des balises pour les domaines existants.

Chaque domaine configuré en mode IAM authentification peut utiliser un espace partagé pour une collaboration en temps quasi réel entre les utilisateurs. Avec un espace partagé, les utilisateurs ont accès à un EFS répertoire Amazon partagé et à une JupyterServerapplication partagée pour l'interface utilisateur, et peuvent co-modifier en temps quasi réel. Le balisage automatique des ressources créées par les espaces partagés permet aux administrateurs de suivre les coûts au niveau du projet. L' JupyterServer interface utilisateur partagée filtre également les ressources telles que les expériences et les entrées de registre des modèles afin que seuls les éléments pertinents pour le projet de machine learning partagé soient affichés. Le schéma suivant fournit une vue d'ensemble des applications privées et des espaces partagés au sein de chaque domaine.




Vue d'ensemble des applications privées et des espaces partagés au sein d'un même domaine

## Configurez des espaces partagés dans votre domaine

Les espaces partagés sont généralement créés pour une entreprise ou un projet de machine learning particulier où les membres d'un même domaine ont besoin d'un accès en temps quasi réel au même stockage de fichiers sous-jacent etIDE. L'utilisateur peut accéder à ses blocs-notes, les lire, les modifier et les partager en temps quasi réel, ce qui lui permet de commencer à itérer avec ses pairs le plus rapidement possible.

Pour créer un espace partagé, vous devez d'abord désigner un rôle d'exécution par défaut qui régira les autorisations de tout utilisateur utilisant l'espace. Au moment de la rédaction de cet article, tous les utilisateurs d'un domaine auront accès à tous les espaces partagés de leur domaine. Reportez-vous à la section Créer un espace partagé pour obtenir la dernière documentation sur l'ajout d'espaces partagés à un domaine existant.

## Configurez votre domaine pour la IAM fédération

Avant de configurer la fédération AWS Identity and Access Management (IAM) pour votre domaine SageMaker AI Studio, vous devez configurer un rôle d'utilisateur de IAM fédération (tel qu'un administrateur de plateforme) dans votre IdP, comme indiqué dans la section Gestion des identités.

Pour obtenir des instructions détaillées sur la configuration d' SageMaker AI Studio avec IAM cette option, reportez-vous à la section Intégration au SageMaker domaine Amazon à l'aide IAM d'Identity Center.

## Configurez votre domaine pour la fédération d'authentification unique (SSO)

Pour utiliser la fédération d'authentification unique (SSO), vous devez l'activer AWS IAM Identity Center dans votre compte de AWS Organizationsgestion dans la même région que celle dans laquelle vous devez exécuter SageMaker AI Studio. Les étapes de configuration du domaine sont similaires aux étapes de IAM fédération, sauf que vous sélectionnez AWS IAM Identity Center(iDC) dans la section Authentification.

Pour obtenir des instructions détaillées, reportez-vous à la section Intégration au SageMaker domaine Amazon à l'aide IAM d'Identity Center.

# SageMaker Profil utilisateur d'AI Studio

Un profil utilisateur représente un utilisateur unique au sein d'un domaine et constitue le principal moyen de référencer une « personne » à des fins de partage, de reporting et d'autres fonctionnalités orientées vers l'utilisateur. Cette entité est créée lorsqu'un utilisateur intègre toSageMaker AI Studio. Si un administrateur invite une personne par e-mail ou l'importe depuis iDC, un profil utilisateur est automatiquement créé. Un profil utilisateur est le principal détenteur des paramètres d'un utilisateur individuel et contient une référence au répertoire personnel privé Amazon Elastic File System (AmazonEFS) de l'utilisateur. Nous vous recommandons de créer un profil utilisateur pour chaque utilisateur physique de l'application SageMaker AI Studio. Chaque utilisateur possède son propre répertoire dédié sur AmazonEFS, et les profils utilisateur ne peuvent pas être partagés entre les domaines d'un même compte.

Chaque profil utilisateur partageant le domaine SageMaker AI Studio reçoit des ressources de calcul dédiées (telles que des instances SageMaker AI Amazon Elastic Compute Cloud (AmazonEC2)) pour exécuter des blocs-notes. Les instances de calcul allouées à l'utilisateur 1 sont complètement isolées de celles allouées à l'utilisateur 2. De même, les ressources informatiques allouées aux utilisateurs d'un AWS compte sont complètement distinctes de celles allouées aux utilisateurs d'un autre compte. Chaque utilisateur peut exécuter jusqu'à quatre applications (applications) dans des conteneurs Docker isolés ou des images sur le même type d'instance.

## Application Jupyter Server

Lorsque vous lancez un bloc-notes Amazon SageMaker AI Studio pour un utilisateur en accédant au pré-signé URL ou en vous connectant à l'aide d'AWSIAMiDC, l'application Jupyter Server est lancée dans l'instance gérée par le SageMaker service AI. VPC Chaque utilisateur dispose de sa propre application Jupyter Server dédiée dans une application privée. Par défaut, l'application Jupyter Server pour ordinateurs portables SageMaker AI Studio est exécutée sur une `ml.t3.medium` instance dédiée (réservée en tant que type d'instance système). Le calcul pour cette instance n'est pas facturé au client.

## L'application Jupyter Kernel Gateway

L'application Kernel Gateway peut être créée via l'interface API ou l'interface SageMaker AI Studio, et elle s'exécute sur le type d'instance choisi. Cette application peut être exécutée à l'aide de l'une des images intégrées d' SageMaker AI Studio préconfigurées avec les logiciels de science des données les plus courants et de deep learning tels qu'TensorFlowApache MXNet et PyTorch.

Les utilisateurs peuvent démarrer et exécuter plusieurs noyaux de bloc-notes Jupyter, des sessions de terminal et des consoles interactives dans le même studio. SageMaker image/Kernel Gateway app. Users can also run up to four Kernel Gateway apps or images on the same physical instance—each isolated by its container/image

Pour créer des applications supplémentaires, vous devez utiliser un autre type d'instance. Un profil utilisateur ne peut avoir qu'une seule instance en cours d'exécution, quel que soit le type d'instance. Par exemple, un utilisateur peut exécuter à la fois un simple bloc-notes utilisant l'image de science des données intégrée à SageMaker AI Studio et un autre bloc-notes utilisant l' TensorFlow image intégrée, sur la même instance. Les utilisateurs sont facturés en fonction de la durée d'exécution de l'instance. Pour éviter des coûts lorsque l'utilisateur n'exécute pas activement SageMaker AI Studio, il doit arrêter l'instance. Pour plus d'informations, reportez-vous à la section Arrêter et mettre à jour les applications Studio.

Chaque fois que vous arrêtez et rouvrez une application Kernel Gateway depuis l'interface SageMaker AI Studio, cette application est démarrée sur une nouvelle instance. Cela signifie que l'installation du package n'est pas maintenue lors des redémarrages de la même application. De même, si un utilisateur change le type d'instance sur un bloc-notes, les packages installés et les variables de session sont perdus. Cependant, vous pouvez utiliser des fonctionnalités telles que l'ajout de votre propre image et des scripts de cycle de vie pour transférer les packages de l'utilisateur dans SageMaker AI Studio et les conserver via des changements d'instance et le lancement de nouvelles instances.

## Volume Amazon Elastic File System

Lorsqu'un domaine est créé, un seul volume Amazon Elastic File System (AmazonEFS) est créé pour être utilisé par tous les utilisateurs du domaine. Chaque profil utilisateur reçoit un répertoire personnel privé dans le EFS volume Amazon pour stocker les blocs-notes, les GitHub référentiels et les fichiers de données de l'utilisateur. Chaque espace d'un domaine reçoit un répertoire privé au sein du EFS volume Amazon auquel plusieurs profils d'utilisateurs peuvent accéder. L'accès aux dossiers est séparé par utilisateur, par le biais des autorisations du système de fichiers. SageMaker AI Studio crée un identifiant utilisateur unique global pour chaque profil ou espace utilisateur, et l'applique en tant qu'interface de système d'exploitation portable (POSIX) pour accéder user/group ID for the user’s home directory on EFS, which prevents other users/spaces à ses données.

## Sauvegarde et restauration

Un EFS volume existant ne peut pas être rattaché à un nouveau domaine SageMaker AI. Dans un environnement de production, assurez-vous que le EFS volume Amazon est sauvegardé (sur un autre EFS volume ou sur Amazon Simple Storage Service (Amazon S3)). Si un EFS volume est supprimé accidentellement, l'administrateur doit démonter et recréer le domaine SageMaker AI Studio. Procédez comme suit :

Sauvegardez la liste des profils utilisateur, des espaces et de EFS l'utilisateur associé IDs (UIDs) via les `DescribeSpace` API appels `ListUserProfiles` `DescribeUserProfileList Spaces`,, et.

1. Créez un nouveau domaine SageMaker AI Studio.
2. Créez les profils utilisateur et les espaces.
3. Pour chaque profil utilisateur, copiez les fichiers de la sauvegarde sur EFS /Amazon S3.
4. Supprimez éventuellement toutes les applications et tous les profils utilisateur de l'ancien domaine SageMaker AI Studio.

Pour des instructions détaillées, reportez-vous à la section annexe relative à la sauvegarde et à la restauration du domaine SageMaker AI Studio.

> **Note**
>
> Cela peut également être réalisé en `LifecycleConfigurations` sauvegardant les données depuis et vers S3 chaque fois qu'un utilisateur démarre son application.

## EBSVolume d'Amazon

Un volume de stockage Amazon Elastic Block Store (AmazonEBS) est également associé à chaque instance d' SageMaker AI Studio Notebook. Il est utilisé comme volume racine du conteneur ou de l'image exécutée sur l'instance. Tant que le EFS stockage Amazon est persistant, le EBS volume Amazon attaché au conteneur est temporaire. Les données stockées localement sur Amazon EBS Volume ne seront pas conservées si le client supprime l'application.

# Sécurisation de l'accès aux documents pré-signés URL

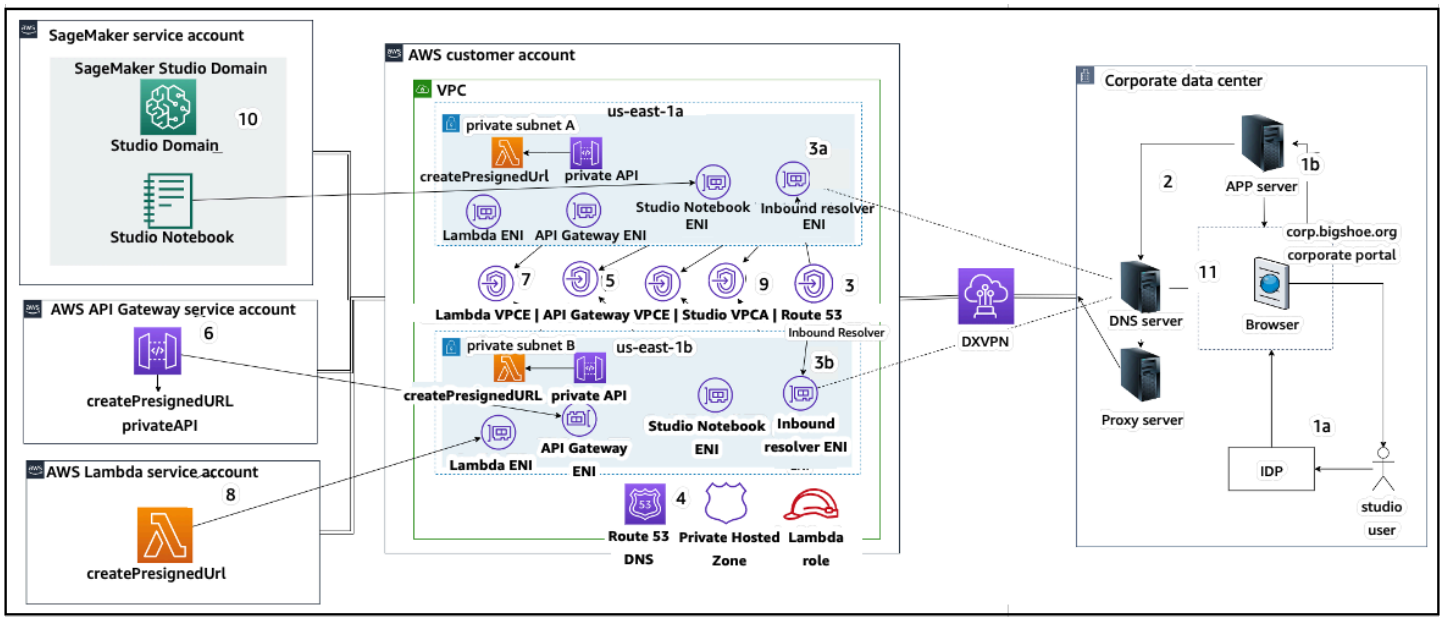
Lorsqu'un utilisateur d' SageMaker AI Studio ouvre le lien du bloc-notes, SageMaker AI Studio valide la IAM politique de l'utilisateur fédéré pour autoriser l'accès, puis génère et résout le lien pré-signé URL pour l'utilisateur. Comme la console SageMaker AI s'exécute sur un domaine Internet, celui-ci généré et pré-signé URL est visible dans la session du navigateur. Cela constitue un vecteur de menace indésirable pour le vol de données et l'accès aux données des clients lorsque les contrôles d'accès appropriés ne sont pas appliqués.

Studio prend en charge plusieurs méthodes pour renforcer les contrôles d'accès contre le vol de URL données pré-signées :

- Validation de l'adresse IP du client en utilisant la condition IAM de politique `aws:sourceIp`
- VPCValidation du client à l'aide de la IAM condition `aws:sourceVpc`
- Validation du VPC terminal client à l'aide de la condition IAM de politique `aws:sourceVpce`

Lorsque vous accédez aux blocs-notes SageMaker AI Studio depuis la console SageMaker AI, la seule option disponible consiste à utiliser la validation de l'adresse IP du client avec la condition IAM `aws:sourceIp` de politique. Cependant, vous pouvez utiliser des produits de routage du trafic par navigateur tels que Zscaler pour garantir l'évolutivité et la conformité de l'accès Internet de votre personnel. Ces produits de routage du trafic génèrent leur propre adresse IP source, dont la plage d'adresses IP n'est pas contrôlée par l'entreprise cliente. Il est donc impossible pour ces entreprises clientes d'utiliser `aws:sourceIp` cette condition.

Pour utiliser la validation du point de VPC terminaison client à l'aide de la condition de IAM politique`aws:sourceVpce`, la création d'un point de terminaison pré-signé URL doit provenir du même client VPC où SageMaker AI Studio est déployé, et la résolution des URL besoins pré-signés doit se faire via un point de VPC terminaison SageMaker AI Studio sur le client. VPC Cette résolution du pré-signé URL pendant le temps d'accès pour les utilisateurs du réseau d'entreprise peut être réalisée à l'aide de règles de DNS transfert (à la fois dans Zscaler et dans l'entrepriseDNS), puis sur le point de VPC terminaison du client à l'aide d'un résolveur entrant Amazon Route 53, comme illustré dans l'architecture suivante :




Accès à Studio pré-signé URL avec VPC Endpoint via le réseau d'entreprise

Pour step-by-step obtenir des conseils sur la configuration de l'architecture précédente, reportez-vous à la section Secure Amazon SageMaker AI Studio présignée, URLs partie 1 : infrastructure de base.

# SageMaker Quotas et limites de domaines AI

- SageMaker La SSO fédération de domaines AI Studio n'est prise en charge que dans la région, sur tous les comptes membres de l' AWS organisation où AWS Identity Center est approvisionné.
- Les espaces partagés ne sont actuellement pas pris en charge avec les domaines configurés avec AWS Identity Center.
- VPCet la configuration du sous-réseau ne peut pas être modifiée après la création du domaine. Vous pouvez toutefois créer un nouveau domaine avec une configuration VPC de sous-réseau différente.
- L'accès au domaine ne peut pas être commuté entre les SSO modes IAM et une fois le domaine créé. Vous pouvez créer un nouveau domaine avec un mode d'authentification différent.
- Il existe une limite de quatre applications de passerelle de noyau par type d'instance lancées pour chaque utilisateur.
- Chaque utilisateur ne peut lancer qu'une seule instance de chaque type d'instance.
- Les ressources consommées au sein d'un domaine sont limitées, telles que le nombre d'instances lancées par type d'instance et le nombre de profils utilisateur pouvant être créés. Reportez-vous à la page des quotas de service pour obtenir la liste complète des limites de service.

- Les clients peuvent soumettre un dossier de support d'entreprise avec une justification commerciale pour augmenter les limites de ressources par défaut, telles que le nombre de domaines ou les profils d'utilisateurs, sous réserve de garanties au niveau du compte.
- La limite stricte du nombre d'applications simultanées par compte est de 2 500 applications. Les limites des domaines et des profils utilisateurs dépendent de cette limite stricte. Par exemple, un compte peut avoir un seul domaine avec 1 000 profils d'utilisateurs, ou 20 domaines avec 50 profils d'utilisateur chacun.

# Gestion des identités

Cette section explique comment les utilisateurs du personnel d'un annuaire d'entreprise se fédérent dans SageMaker AI Studio Comptes AWS et y accèdent. Tout d'abord, nous allons décrire brièvement comment les utilisateurs, les groupes et les rôles sont mappés, ainsi que le fonctionnement de la fédération d'utilisateurs.

## Utilisateurs, groupes et rôles

Dans AWS, les autorisations relatives aux ressources sont gérées à l'aide d'utilisateurs, de groupes et de rôles. Les clients peuvent gérer leurs utilisateurs et leurs groupes via IAM ou dans un annuaire d'entreprise tel qu'Active Directory (AD), activé via un IdP externe tel qu'Okta, qui leur permet d'authentifier les utilisateurs auprès de diverses applications exécutées dans le cloud et sur site.

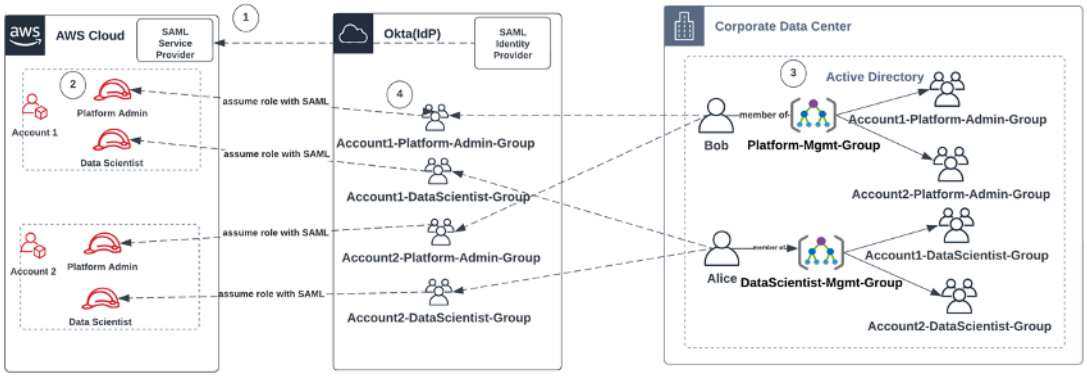
Comme indiqué dans la section Gestion des identités du pilier de AWS sécurité, il est recommandé de gérer les identités de vos utilisateurs dans un IdP central, car cela permet de s'intégrer facilement à vos processus RH principaux et de gérer l'accès aux utilisateurs de votre personnel.

IdPs tels qu'Okta, permettent aux utilisateurs finaux de s'authentifier auprès d'un ou de plusieurs rôles Comptes AWS et d'accéder à des rôles spécifiques à l'aide du langage de balisage SSO d'assermentation de sécurité (). SAML Les administrateurs d'IdP ont la possibilité de télécharger des rôles depuis l' Comptes AWS IdP et de les attribuer aux utilisateurs. Lorsqu'ils se connectent à AWS, les utilisateurs finaux voient apparaître un AWS écran qui affiche une liste des AWS rôles qui leur ont été assignés dans un ou plusieurs rôles Comptes AWS. Ils peuvent sélectionner le rôle à assumer pour la connexion, qui définit leurs autorisations pour la durée de cette session authentifiée.

Un groupe doit exister dans IdP pour chaque combinaison de comptes et de rôles spécifique à laquelle vous souhaitez donner accès. Vous pouvez considérer ces groupes comme des groupes AWS spécifiques à un rôle. Tout utilisateur membre de ces groupes spécifiques à un rôle se voit octroyer un droit unique : l'accès à un rôle spécifique dans un rôle spécifique Compte AWS. Toutefois, ce processus d'autorisation unique ne permet pas de gérer l'accès des utilisateurs en attribuant à chaque utilisateur des groupes de AWS rôles spécifiques. Pour simplifier l'administration, nous vous recommandons également de créer un certain nombre de groupes pour tous les groupes d'utilisateurs distincts de votre organisation qui nécessitent différents ensembles de AWS droits.

Pour illustrer la configuration centrale de l'IdP, imaginons une entreprise dotée d'une configuration AD, dans laquelle les utilisateurs et les groupes sont synchronisés avec le répertoire IdP. Dans AWS,

ces groupes AD sont mappés à IAM des rôles. Les principales étapes du flux de travail sont les suivantes :




Flux de travail pour l'intégration des utilisateurs, des groupes et IAM des rôles AD

1. Dans AWS, configurez SAML l'intégration pour chacun d'entre vous Comptes AWS avec votre IdP.
2. Dans AWS, configurez des rôles dans chacun d'eux Compte AWS et synchronisez-les avec IdP.
3. Dans le système AD d'entreprise :
   a. Créez un groupe AD pour chaque rôle de compte et synchronisez-le avec IdP (par exemple, `Account1-Platform-Admin-Group` (alias AWS Role Group)).
   b. Créez un groupe de gestion à chaque niveau de personnalité (par exemple,`Platform-Mgmt-Group`) et assignez des groupes de AWS rôles en tant que membres.
   c. Affectez des utilisateurs à ce groupe de gestion pour autoriser l'accès aux Compte AWS rôles.
4. Dans IdP, associez des groupes de AWS rôles (tels que`Account1-Platform-Admin-Group`) à Compte AWS des rôles (tels que Platform Admin dans Account1).
5. Lorsque la data scientist Alice se connecte à Idp, une interface utilisateur de l'application AWS Federation lui est présentée, avec deux options parmi lesquelles choisir : « Account 1 Data Scientist » et « Account 2 Data Scientist ».
6. Alice choisit l'option « Account 1 Data Scientist », et ils sont connectés à leur application autorisée dans AWS Account 1 (SageMaker AI Console).

Pour obtenir des instructions détaillées sur la configuration de la fédération de SAML comptes, reportez-vous à la section Comment configurer la fédération de AWS comptes SAML 2.0 d'Okta.

# Fédération d'utilisateurs

L'authentification pour SageMaker AI Studio peut être effectuée à l'aide d'iDC IAM ou d'IAMiDC. Si les utilisateurs sont gérés viaIAM, ils peuvent choisir le IAM mode. Si l'entreprise utilise un IdP externe, elle peut soit fédérer, soit via IAM IdC. IAM Notez que le mode d'authentification ne peut pas être mis à jour pour un domaine SageMaker AI Studio existant. Il est donc essentiel de prendre une décision avant de créer un domaine SageMaker AI Studio de production.

Si SageMaker AI Studio est configuré en IAM mode, les utilisateurs d' SageMaker AI Studio accèdent à l'application via un système pré-signé URL qui connecte automatiquement un utilisateur à l'application SageMaker AI Studio lorsqu'il y accède via un navigateur.

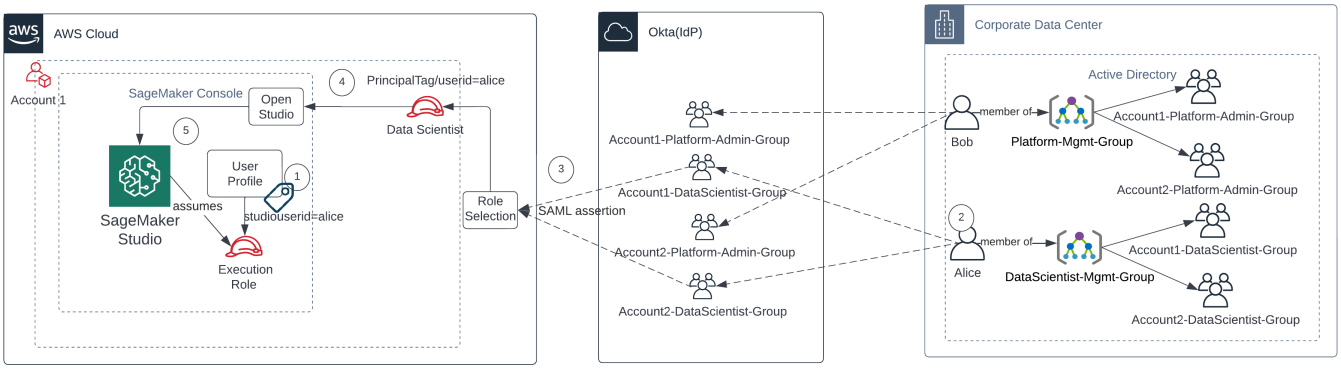
## Utilisateurs IAM

Pour IAM les utilisateurs, l'administrateur crée des profils utilisateur SageMaker AI Studio pour chaque utilisateur et associe le profil utilisateur à un IAM rôle qui autorise les actions nécessaires que l'utilisateur doit effectuer depuis Studio. Pour empêcher un AWS utilisateur d'accéder uniquement à son profil utilisateur SageMaker AI Studio, l'administrateur doit étiqueter le profil utilisateur SageMaker AI Studio et associer à l'utilisateur une IAM politique lui permettant d'accéder uniquement si la valeur de la balise est identique au nom AWS d'utilisateur. La déclaration de politique se présente comme suit :

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AmazonSageMakerPresignedUrlPolicy",
            "Effect": "Allow",
            "Action": [
                "sagemaker:CreatePresignedDomainUrl"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {
                    "sagemaker:ResourceTag/studiouserid": "${aws:username}"
                }
            }
        }
    ]
}
```

# AWS IAMou fédération de comptes

La méthode Compte AWS de fédération permet aux clients de se fédérer dans la console SageMaker AI à partir de leur SAML IdP, tel qu'Okta. Pour empêcher les utilisateurs d'accéder uniquement à leur profil utilisateur, l'administrateur doit baliser le profil utilisateur d' SageMaker AI Studio, ajouter `PrincipalTags` l'IdP et les définir comme des balises transitives. Le schéma suivant montre comment l'utilisateur fédéré (Data Scientist Alice) est autorisé à accéder à son propre profil utilisateur SageMaker AI Studio.




Accès à SageMaker AI Studio en IAM mode fédération

1. Le profil utilisateur d'Alice SageMaker AI Studio est balisé avec son ID utilisateur et associé au rôle d'exécution.
2. Alice s'authentifie auprès de l'IdP (Okta).
3. L'IdP authentifie Alice et publie une SAML assertion avec les deux rôles (Data Scientist pour les comptes 1 et 2) dont Alice est membre. Alice sélectionne le rôle de data scientist pour le compte 1.
4. Alice est connectée à la console SageMaker AI du compte 1, assumant le rôle de data scientist. Alice ouvre son instance d'application Studio à partir de la liste des instances d'application Studio.
5. La balise principale Alice dans la session de rôle assumé est validée par rapport à la balise de profil utilisateur de l'instance d'application SageMaker AI Studio sélectionnée. Si la balise de profil est valide, l'instance de l'application SageMaker AI Studio est lancée en assumant le rôle d'exécution.

Si vous souhaitez automatiser la création de rôles et de politiques d'exécution de l' SageMaker IA dans le cadre de l'intégration des utilisateurs, voici un moyen d'y parvenir :

1. Configurez un groupe AD, par exemple `SageMaker AI-Account1-Group` au niveau de chaque compte et du domaine Studio.
2. Ajoutez SageMaker AI-Account1-Group à l'adhésion au groupe de l'utilisateur lorsque vous devez intégrer un utilisateur à AI Studio. SageMaker

Configurez un processus d'automatisation qui écoute l'événement `SageMaker AI-Account1-Group` d'adhésion et utilisez-le AWS APIs pour créer le rôle, les politiques, les balises et le profil utilisateur d' SageMaker AI Studio en fonction de leur appartenance au groupe AD. Associez le rôle au profil utilisateur. Pour un exemple de politique, reportez-vous àEmpêcher les utilisateurs d' SageMaker AI Studio d'accéder à d'autres profils utilisateur.

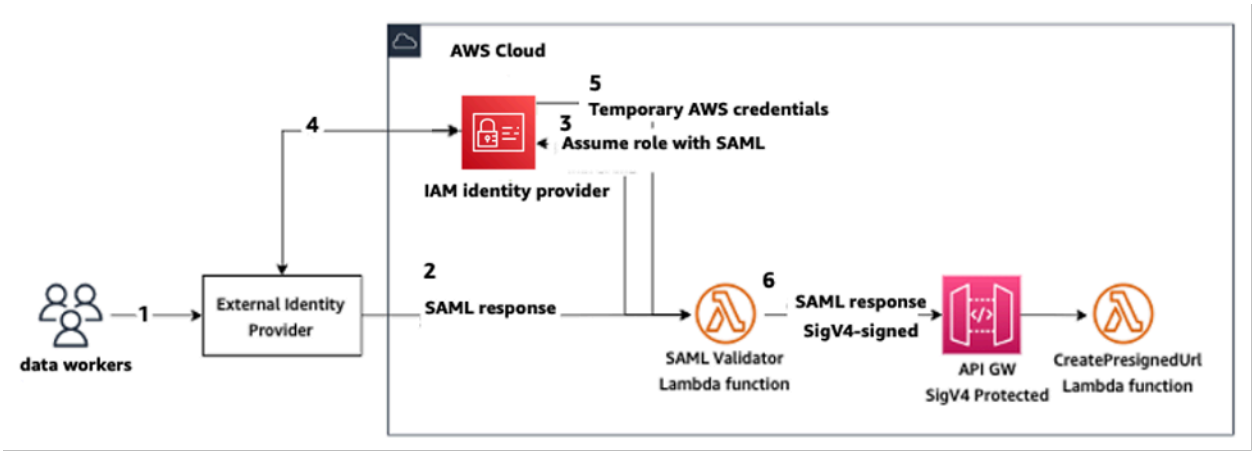
## SAMLauthentification à l'aide de AWS Lambda

En IAM mode, les utilisateurs peuvent également être authentifiés dans SageMaker AI Studio à l'aide d'SAMLassertions. Dans cette architecture, le client dispose d'un IdP existant, dans lequel il peut créer une SAML application permettant aux utilisateurs d'accéder à Studio (au lieu de l'application AWS Identity Federation). L'IdP du client est ajouté à. IAM Une AWS Lambda fonction permet de valider l'SAMLassertion en utilisant IAM etSTS, puis en invoquant directement une API passerelle ou une fonction Lambda, pour créer le domaine pré-signé. URL

L'avantage de cette solution est que la fonction Lambda peut personnaliser la logique d'accès à SageMaker AI Studio. Par exemple :

- Créez automatiquement un profil utilisateur s'il n'en existe pas.
- Attachez ou supprimez des rôles ou des documents de politique au rôle d'exécution d' SageMaker AI Studio en analysant les SAML attributs.
- Personnalisez le profil utilisateur en ajoutant la configuration du cycle de vie (LCC) et en ajoutant des balises.

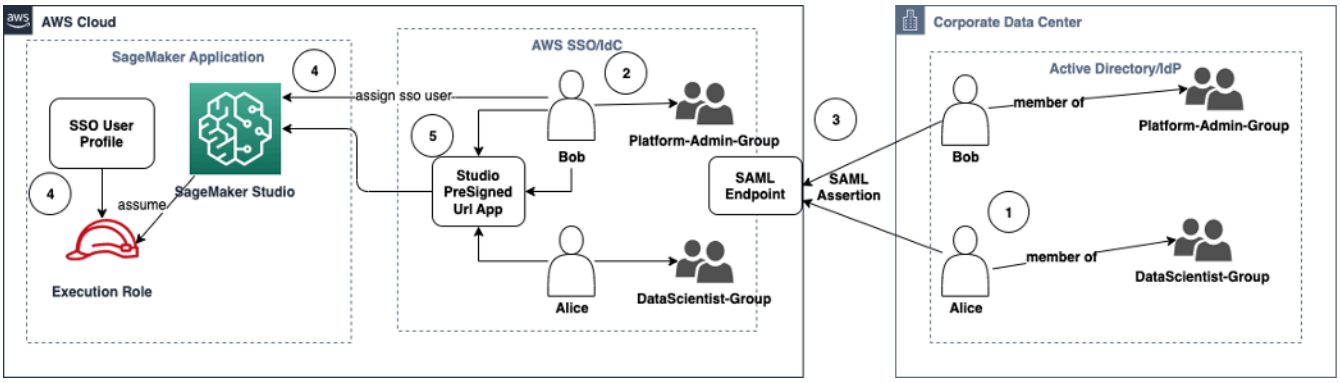
En résumé, cette solution exposera SageMaker AI Studio en tant qu'application SAML2 .0 avec une logique personnalisée pour l'authentification et l'autorisation. Reportez-vous à la section annexe Accès au SageMaker studio à l'aide de l'SAMLassertion pour les détails de mise en œuvre.




Accès à SageMaker AI Studio à l'aide d'une SAML application personnalisée

# AWSIAMFédération iDC

La méthode de fédération iDC permet aux clients de se fédérer directement dans l'application SageMaker AI Studio à partir de leur SAML IdP (tel qu'Okta). Le schéma suivant montre comment l'utilisateur fédéré est autorisé à accéder à sa propre instance SageMaker AI Studio.




Accès à SageMaker AI Studio en mode IAM iDC

1. Dans l'AD d'entreprise, l'utilisateur est membre de groupes AD tels que le groupe Platform Admin et le groupe Data Scientist.
2. L'utilisateur AD et les groupes AD du fournisseur d'identité (IdP) sont synchronisés avec AWS IAM Identity Center et disponibles en tant qu'utilisateurs et groupes d'authentification unique pour les attributions, respectivement.
3. L'IdP publie une SAML assertion sur le point de terminaison AWS SAML iDC.
4. Dans SageMaker AI Studio, l'utilisateur iDC est affecté à l'application SageMaker Studio. Cette attribution peut être effectuée à l'aide d'iDC Group et SageMaker AI Studio s'appliquera à chaque

niveau d'utilisateur iDC. Lorsque cette attribution est créée, SageMaker AI Studio crée un profil utilisateur iDC et attache le rôle d'exécution du domaine.

5. L'utilisateur accède à l'application SageMaker AI Studio à l'aide de l'application sécurisée présignée URL hébergée en tant qu'application cloud depuis l'iDC. SageMaker AI Studio assume le rôle d'exécution associé à leur profil utilisateur iDC.

## Conseils d'authentification de domaine

Voici quelques points à prendre en compte lors du choix du mode d'authentification d'un domaine :

1. Si vous souhaitez que vos utilisateurs n'accèdent pas directement à l'interface utilisateur d' SageMaker AI Studio AWS Management Console et ne la consultent pas directement, utilisez le mode d'authentification unique avec AWS IAM iDC.
2. Si vous souhaitez que vos utilisateurs n'accèdent pas à l'interface utilisateur d' SageMaker AI Studio AWS Management Console et ne la consultent pas directement en IAM mode, vous pouvez le faire en utilisant une fonction Lambda dans le backend URL pour générer un profil utilisateur présigné et en les redirigeant vers l'interface utilisateur d'AI Studio. SageMaker
3. En mode iDC, chaque utilisateur est mappé à un profil utilisateur unique.
4. Le rôle d'exécution par défaut est automatiquement attribué à tous les profils utilisateur en mode iDC. Si vous souhaitez que différents rôles d'exécution soient assignés à vos utilisateurs, vous devez mettre à jour les profils utilisateurs à l'aide du UpdateUserProfileAPI.
5. Si vous souhaitez restreindre l'accès à l'interface utilisateur d' SageMaker AI Studio en IAM mode (à l'aide du pré-signé généréURL) à un VPC point de terminaison, sans passer par Internet, vous pouvez utiliser un résolveur personnaliséDNS. Reportez-vous au billet de blog présigné par Secure Amazon SageMaker AI Studio, URLs partie 1 : Infrastructure fondamentale.

# Gestion des autorisations

Cette section décrit les meilleures pratiques pour configurer les IAM rôles, les politiques et les garde-fous couramment utilisés pour le provisionnement et l'exploitation du domaine SageMaker AI Studio.

## Rôles et stratégies IAM

La meilleure pratique consiste à identifier d'abord les personnes et les applications pertinentes, appelées « responsables » impliqués dans le cycle de vie du machine learning, et les AWS autorisations que vous devez leur accorder. SageMaker L'IA étant un service géré, vous devez également prendre en compte les principes de service, qui sont des AWS services qui peuvent API passer des appels au nom d'un utilisateur. Le schéma suivant illustre les différents IAM rôles que vous souhaiterez peut-être créer, correspondant aux différents personnages de l'organisation.

SageMaker IAMRôles de l'IA

Ces rôles sont décrits en détail, avec quelques exemples spécifiques dont IAMpermissions ils auront besoin.

- Rôle utilisateur d'administrateur ML : il s'agit d'un directeur qui fournit l'environnement aux scientifiques des données en créant des domaines de studio et des profils utilisateur (`sagemaker:CreateDomain`,`sagemaker:CreateUserProfile`), en créant AWS Key Management Service des clés pour les utilisateurs, en créant des compartiments S3 pour les scientifiques des données et en créant des ECR référentiels Amazon pour héberger des conteneurs.AWS KMS Ils peuvent également définir des configurations par défaut et des scripts de cycle de vie pour les utilisateurs, créer et joindre des images personnalisées au domaine SageMaker AI Studio, et fournir des produits Service Catalog tels que des projets personnalisés et des EMR modèles Amazon.

  Comme ce directeur n'exécutera pas de tâches de formation, par exemple, il n'a pas besoin d'autorisations pour lancer des tâches de formation ou de traitement liées à l' SageMaker IA. S'ils utilisent l'infrastructure comme modèles de code, tels que CloudFormation Terraform,

pour approvisionner des domaines et des utilisateurs, ce rôle sera assumé par le service de provisionnement pour créer les ressources au nom de l'administrateur. Ce rôle peut avoir un accès en lecture seule à l' SageMaker IA à l'aide du. AWS Management Console

Ce rôle d'utilisateur aura également besoin de certaines EC2 autorisations pour lancer le domaine dans un espace privéVPC, d'KMSautorisations pour chiffrer le EFS volume, ainsi que d'autorisations pour créer un rôle lié à un service pour Studio (`iam:CreateServiceLinkedRole`). Nous décrirons ces autorisations détaillées plus loin dans le document.

- Rôle d'utilisateur du data scientist : ce principe est celui de l'utilisateur qui se connecte à SageMaker AI Studio, explore les données, crée des tâches et des pipelines de traitement et de formation, etc. L'autorisation principale dont l'utilisateur a besoin est l'autorisation de lancer SageMaker AI Studio, et le reste des politiques peut être géré par le rôle de service d'exécution de l' SageMaker IA.
- SageMaker Rôle du service d'exécution de l' SageMaker IA — L'IA étant un service géré, elle lance des tâches pour le compte d'un utilisateur. Ce rôle est souvent le plus large en termes d'autorisations autorisées, car de nombreux clients choisissent d'utiliser un seul rôle d'exécution pour exécuter des tâches de formation, des tâches de traitement ou des tâches d'hébergement de modèles. Bien qu'il s'agisse d'un moyen simple de démarrer, les clients évoluant au fil de leur parcours, ils divisent souvent le rôle d'exécution du bloc-notes en rôles distincts pour différentes API actions, en particulier lorsqu'ils exécutent ces tâches dans des environnements déployés.

  Vous associez un rôle au domaine SageMaker AI Studio lors de sa création. Toutefois, comme les clients peuvent avoir besoin de la flexibilité d'avoir différents rôles associés aux différents profils utilisateur du domaine (par exemple, en fonction de leur fonction), vous pouvez également associer un IAM rôle distinct à chaque profil utilisateur. Nous vous recommandons de mapper un seul utilisateur physique à un profil utilisateur unique. Si vous n'associez aucun rôle à un profil utilisateur lors de sa création, le comportement par défaut consiste également à associer le rôle d'exécution du SageMaker AIStudio domaine au profil utilisateur.

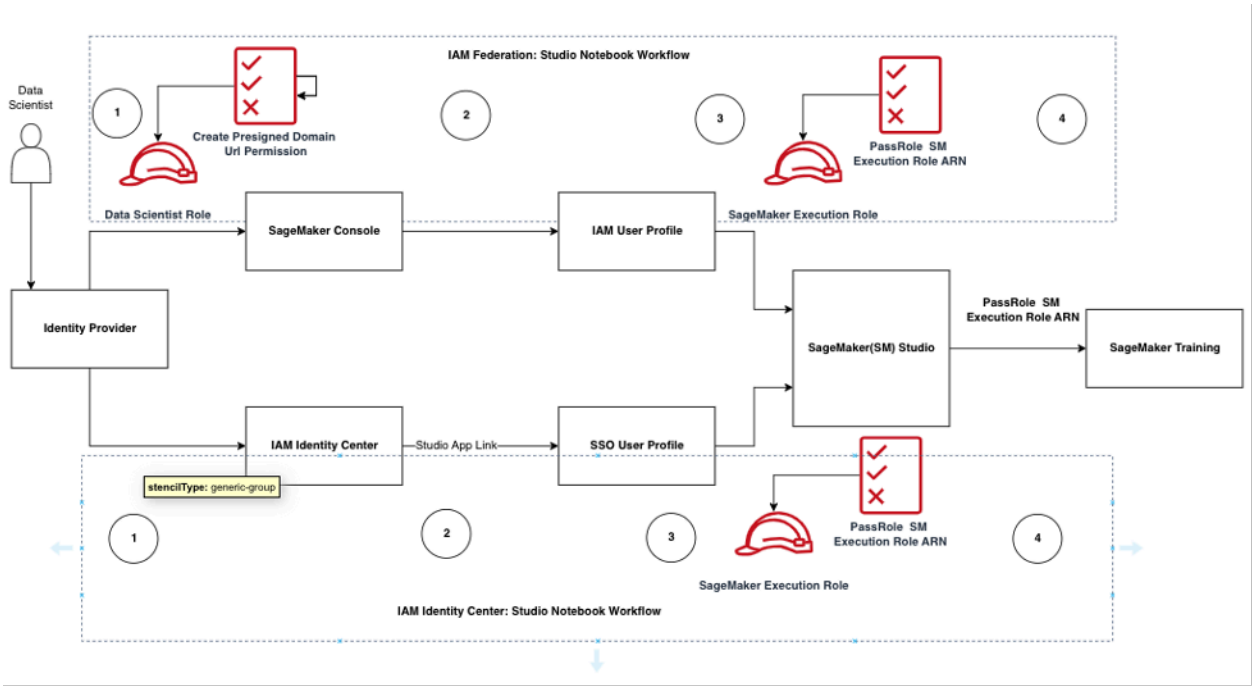
  Dans les cas où plusieurs scientifiques des données et ingénieurs du ML travaillent ensemble sur un projet et ont besoin d'un modèle d'autorisation partagé pour accéder aux ressources, nous vous recommandons de créer un rôle d'exécution de service d' SageMaker IA au niveau de l'équipe pour partager les IAM autorisations entre les membres de votre équipe. Dans les cas où vous devez verrouiller les autorisations à chaque niveau d'utilisateur, vous pouvez créer un rôle d'exécution de service d' SageMaker IA individuel au niveau de l'utilisateur ; vous devez toutefois tenir compte de vos limites de service.

# SageMaker Flux de travail d'autorisation d'AI Studio Notebook

Cette section explique comment l'autorisation SageMaker AI Studio Notebook fonctionne pour les différentes activités que le data scientist doit effectuer pour créer et entraîner le modèle directement à partir de l' SageMaker AI Studio Notebook. Le domaine SageMaker AI prend en charge deux modes d'autorisation :

- IAMfédération
- IAMCentre d'identité

Ensuite, ce paper explique le flux de travail d'autorisation du Data Scientist pour chacun de ces modes.




Flux de travail d'authentification et d'autorisation pour les utilisateurs de Studio

## IAMFédération : flux de travail de SageMaker Studio Notebook

1. Un Data Scientist s'authentifie auprès de son fournisseur d'identité d'entreprise et assume le rôle d'utilisateur Data Scientist (le rôle de fédération d'utilisateurs) dans la console SageMaker AI. Ce rôle de fédération est `iam:PassRole` API autorisé sur le rôle d'exécution SageMaker AI à transmettre le rôle Amazon Resource Name (ARN) à SageMaker Studio.

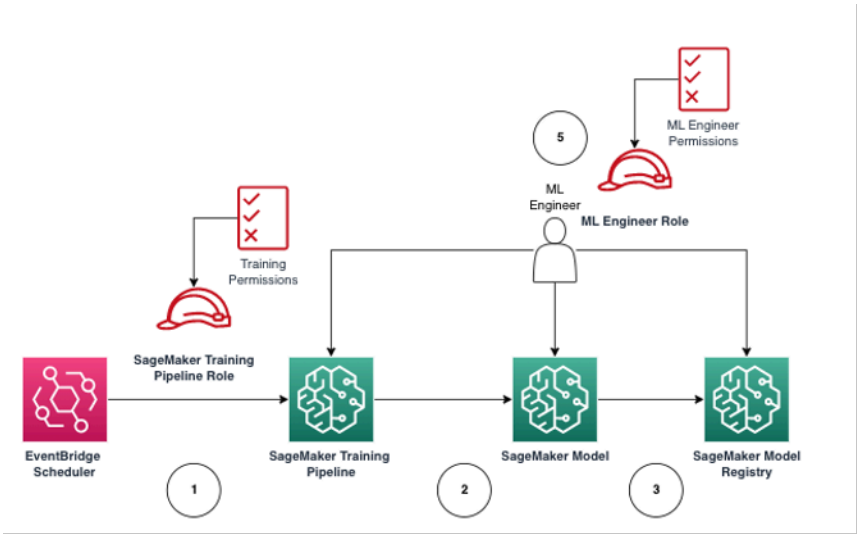
2. Le Data Scientist sélectionne le lien Open Studio dans son profil IAM utilisateur Studio associé au rôle d'exécution de l' SageMaker IA.
3. Le IDE service SageMaker Studio est lancé, en supposant les autorisations de rôle d' SageMaker exécution du profil utilisateur. Ce rôle est `iam:PassRole` API autorisé sur le rôle d'exécution de l' SageMaker IA à transmettre le rôle ARN au service de formation à l' SageMaker IA.
4. Lorsque le Data Scientist lance la tâche de formation dans le ou les nœuds de calcul distants, le rôle d'exécution de l' SageMaker IA ARN est transmis au service de formation de l' SageMaker IA. Cela permet de créer une nouvelle session de rôle ARN et d'exécuter la tâche de formation. Si vous devez définir davantage l'autorisation pour un poste de formation, vous pouvez créer un rôle spécifique à la formation et transmettre ce rôle ARN lorsque vous appelez trainingAPI.

### IAMIdentity Center : flux de travail d' SageMaker AI Studio Notebook

1. Le data scientist s'authentifie auprès de son fournisseur d'identité d'entreprise et clique sur AWS IAM Identity Center. Le Data Scientist reçoit le portail Identity Center pour l'utilisateur.
2. Le Data Scientist clique sur le lien de l'application SageMaker AI Studio créé à partir de son profil utilisateur iDC, qui est associé au rôle d'exécution de l' SageMaker IA.
3. Le IDE service SageMaker AI Studio est lancé, en supposant les autorisations du rôle d'exécution SageMaker AI du profil utilisateur. Ce rôle est `iam:PassRole` API autorisé sur le rôle d'exécution de l' SageMaker IA à transmettre le rôle ARN au service de formation à l' SageMaker IA.
4. Lorsque le Data Scientist lance la tâche de formation dans un ou plusieurs nœuds de calcul distants, le rôle d'exécution de l' SageMaker IA ARN est transmis au service de formation de l' SageMaker IA. Le rôle d'exécution ARN crée ainsi une nouvelle session de rôle et exécute la tâche de formation. ARN Si vous devez limiter davantage l'autorisation pour les tâches de formation, vous pouvez créer un rôle spécifique à la formation et transmettre ce rôle ARN lorsque vous appelez la formation. API

## Environnement déployé : flux de formation SageMaker basé sur l'IA

Dans les environnements déployés tels que les tests et la production de systèmes, les tâches sont exécutées via un planificateur automatique et des déclencheurs d'événements, et l'accès humain à ces environnements est restreint depuis les ordinateurs portables SageMaker AI Studio. Cette section explique comment IAM les rôles fonctionnent avec le pipeline de formation à l' SageMaker IA dans l'environnement déployé.




SageMaker Flux de formation basé sur l'IA dans un environnement de production géré

1. Amazon EventBridge Scheduler déclenche le job SageMaker AI Training Pipeline.
2. Le job du pipeline de formation SageMaker SageMaker AI assume le rôle du pipeline de formation de l'IA pour entraîner le modèle.
3. Le modèle d' SageMaker IA entraîné est enregistré dans le registre des modèles d' SageMaker IA.
4. Un ingénieur ML assume le rôle d'utilisateur de l'ingénieur ML pour gérer le pipeline de formation et le modèle d' SageMaker IA.

# Autorisations relatives aux données

La capacité des utilisateurs d' SageMaker AI Studio à accéder à n'importe quelle source de données est régie par les autorisations associées à leur rôle IAM d'exécution SageMaker AI. Les politiques associées peuvent les autoriser à lire, écrire ou supprimer des données de certains compartiments ou préfixes Amazon S3, et à se connecter aux bases de données AmazonRDS.

## Accès aux AWS Lake Formation données

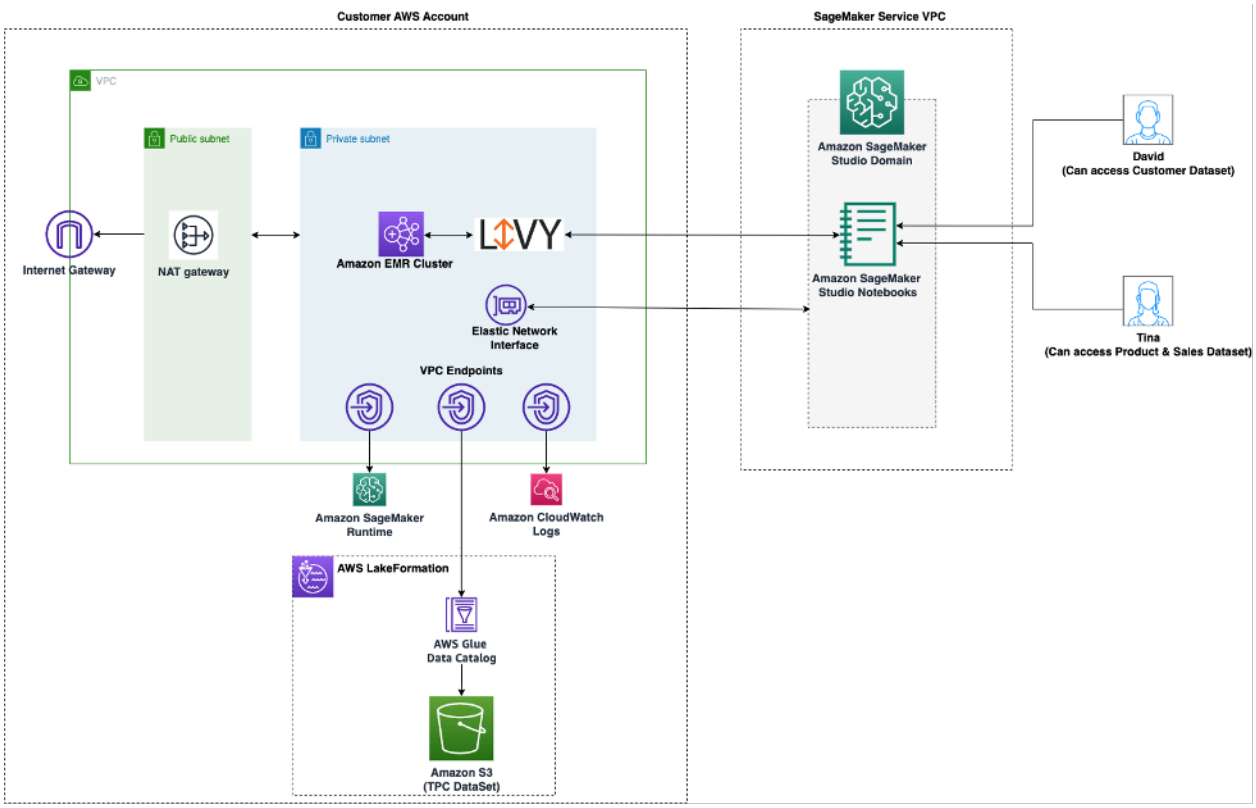
De nombreuses entreprises ont commencé à utiliser des lacs de données régis AWS Lake Formationpour permettre à leurs utilisateurs d'accéder aux données de manière précise. À titre d'exemple de telles données gouvernées, les administrateurs peuvent masquer des colonnes sensibles pour certains utilisateurs tout en autorisant les requêtes de la même table sous-jacente.

Pour utiliser Lake Formation depuis SageMaker AI Studio, les administrateurs peuvent enregistrer les rôles IAM d'exécution de l' SageMaker IA en tant que`DataLakePrincipals`. Pour plus d'informations, reportez-vous à la section Lake Formation Permissions Reference. Une fois autorisé, il existe trois méthodes principales pour accéder aux données gouvernées et les écrire à partir d' SageMaker AI Studio :

1. À partir d'un bloc-notes SageMaker AI Studio, les utilisateurs peuvent utiliser des moteurs de requêtes tels qu'Amazon Athena ou des bibliothèques basées sur boto3 pour extraire des données directement vers le bloc-notes. The AWSSDKfor Pandas (anciennement connue sous le nom de awswrangler) est une bibliothèque populaire. Voici un exemple de code pour montrer à quel point cela peut être simple :

```
transaction_id = wr.lakeformation.start_transaction(read_only=True)
df = wr.lakeformation.read_sql_query(
    sql=f"SELECT * FROM {table};",
    database=database,
    transaction_id=transaction_id
)
```

2. Utilisez la connectivité native d' SageMaker AI Studio à Amazon EMR pour lire et écrire des données à grande échelle. En utilisant les rôles EMR d'exécution Apache Livy et Amazon, SageMaker AI Studio a développé une connectivité native qui vous permet de transmettre votre IAM rôle d'exécution SageMaker AI (ou un autre rôle autorisé) à un EMR cluster Amazon pour l'accès aux données et leur traitement. Reportez-vous à la section Connect to an Amazon EMR Cluster from Studio pour up-to-date obtenir des instructions.




Architecture d'accès aux données gérées par Lake Formation depuis SageMaker Studio

3. Utilisez la connectivité native d' SageMaker AI Studio pour les sessions AWS Glue interactives afin de lire et d'écrire des données à grande échelle. SageMaker Les ordinateurs portables AI Studio ont des noyaux intégrés qui permettent aux utilisateurs d'exécuter des commandes de manière interactive. AWS Glue Cela permet une utilisation évolutive des backends Python, Spark ou Ray qui peuvent lire et écrire des données en toute fluidité à grande échelle à partir de sources de données gouvernées. Les noyaux permettent aux utilisateurs de transmettre leur rôle SageMaker d'exécution ou d'autres IAM rôles autorisés. Reportez-vous à la section Préparation des données à l'aide de sessions AWS Glue interactives pour plus d'informations.

# Rambardes communes

Cette section décrit les garde-fous les plus couramment utilisés pour appliquer la gouvernance à vos ressources ML à l'aide de politiques, de politiques de ressources, de IAM politiques de point de VPC terminaison et de politiques de contrôle des services (). SCPs

## Limitez l'accès au bloc-notes à des instances spécifiques

Cette politique de contrôle des services peut être utilisée pour limiter les types d'instances auxquels les data scientists ont accès lors de la création de blocs-notes Studio. Notez que tout utilisateur aura besoin de l'instance « système » autorisée pour créer l'application Jupyter Server par défaut qui héberge SageMaker AI Studio.

```
{
     "Version": "2012-10-17",
     "Statement": [
         {
             "Sid": "LimitInstanceTypesforNotebooks",
             "Effect": "Deny",
             "Action": [
                 "sagemaker:CreateApp"
             ],
             "Resource": "*",
             "Condition": {
                 "ForAnyValue:StringNotLike": {
                     "sagemaker:InstanceTypes": [
                         "ml.c5.large",
                         "ml.m5.large",
                         "ml.t3.medium",
                         "system"
                     ]
                 }
             }
         }
     ]
 }
```

## Limiter les domaines SageMaker AI Studio non conformes

Pour les domaines SageMaker AI Studio, la politique de contrôle des services suivante peut être utilisée pour obliger le trafic à accéder aux ressources des clients afin qu'elles ne passent pas par l'Internet public, mais plutôt par celui d'un client VPC :

```
{
     "Version": "2012-10-17",
     "Statement": [
         {   "Sid": "LockDownStudioDomain",
             "Effect": "Deny",
```

```
            "Action": [
                "sagemaker:CreateDomain"
            ],
            "Resource": "*",
            "Condition": {
                        "StringNotEquals": {"sagemaker:AppNetworkAccessType":
 "VpcOnly"
                },
                "Null": {
                        "sagemaker:VpcSubnets": "true",
                        "sagemaker:VpcSecurityGroupIds": "true"
                }
            }
        }
    ]
 }
```

## Limitez le lancement d'images SageMaker IA non autorisées

La politique suivante empêche un utilisateur de lancer une image SageMaker AI non autorisée dans son domaine : f

```
{
     "Version": "2012-10-17",
     "Statement": [
         {
             "Action": [
                 "sagemaker:CreateApp"
              ],
             "Effect": "Allow",
             "Resource": "*",
             "Condition": {
                 "ForAllValues:StringNotLike": {
                     "sagemaker:ImageArns":
                         [
                         "arn:aws:sagemaker:*:*:image/{ImageName}"
                         ]
                 }
             }
         }
     ]
 }
```

## Lancez des blocs-notes uniquement via des points de terminaison basés sur SageMaker l'IA VPC

Outre les VPC points de terminaison pour le plan de contrôle SageMaker AI, SageMaker AI prend en charge les VPC points de terminaison permettant aux utilisateurs de se connecter aux blocs-notes SageMaker AI Studio ou aux instances de blocs-notes SageMaker AI. Si vous avez déjà configuré un VPC point de terminaison pour une instance SageMaker AI Studio/Notebook, la clé de IAM condition suivante n'autorisera les connexions aux blocs-notes SageMaker AI Studio que si elles sont établies via le point de terminaison SageMaker AI Studio ou via le SageMaker point de VPC terminaison AI. API

```
{
     "Version": "2012-10-17",
     "Statement": [
         {
             "Sid": "EnableSageMakerStudioAccessviaVPCEndpoint",
             "Effect": "Allow",
             "Action": [
                 "sagemaker:CreatePresignedDomainUrl",
                 "sagemaker:DescribeUserProfile"
             ],
             "Resource": "*",
             "Condition": {
                 "ForAnyValue:StringEquals": {
                     "aws:sourceVpce": [
                         "vpce-111bbccc",
                         "vpce-111bbddd"
                     ]
                 }
             }
         }
     ]
 }
```

## Limitez SageMaker l'accès aux ordinateurs portables AI Studio à une plage d'adresses IP limitée

Les entreprises limitent souvent l'accès à SageMaker AI Studio à certaines plages d'adresses IP d'entreprise autorisées. La IAM politique suivante avec la clé de `SourceIP` condition peut limiter cela.

```
{
     "Version": "2012-10-17",
     "Statement": [
         {
             "Sid": "EnableSageMakerStudioAccess",
             "Effect": "Allow",
             "Action": [
                 "sagemaker:CreatePresignedDomainUrl",
                 "sagemaker:DescribeUserProfile"
             ],
             "Resource": "*",
             "Condition": {
                 "IpAddress": {
                     "aws:SourceIp": [
                         "192.0.2.0/24",
                         "203.0.113.0/24"
                     ]
                 }
             }
         }
     ]
 }
```

## Empêcher les utilisateurs d' SageMaker AI Studio d'accéder à d'autres profils utilisateur

En tant qu'administrateur, lorsque vous créez le profil utilisateur, assurez-vous que le profil est étiqueté avec le nom d'utilisateur SageMaker AI Studio avec la clé de balise`studiouserid`. Le principal (utilisateur ou rôle attaché à l'utilisateur) doit également avoir une étiquette avec la clé `studiouserid` (cette balise peut porter n'importe quel nom et n'est pas limitée à`studiouserid`).

Ensuite, associez la politique suivante au rôle que l'utilisateur assumera lors du lancement d' SageMaker AI Studio.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AmazonSageMakerPresignedUrlPolicy",
            "Effect": "Allow",
            "Action": [
```

```
                "sagemaker:CreatePresignedDomainUrl"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {
                    "sagemaker:ResourceTag/studiouserid": "${aws:PrincipalTag/
studiouserid}"
                }
            }
        }
    ]
}
```

## Appliquer le balisage

Les data scientists doivent utiliser les blocs-notes SageMaker AI Studio pour explorer les données, ainsi que pour créer et entraîner des modèles. L'application de balises aux ordinateurs portables permet de surveiller l'utilisation et de contrôler les coûts, tout en garantissant la propriété et l'auditabilité.

Pour les applications SageMaker AI Studio, assurez-vous que le profil utilisateur est balisé. Les balises sont automatiquement propagées aux applications à partir du profil utilisateur. Pour imposer la création de profils utilisateur à l'aide de balises (prises en charge par CLI etSDK), pensez à ajouter cette politique au rôle d'administrateur :

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "EnforceUserProfileTags",
            "Effect": "Allow",
            "Action": "sagemaker:CreateUserProfile",
            "Resource": "*",
            "Condition": {
                "ForAnyValue:StringEquals": {
                    "aws:TagKeys": [
                        "studiouserid"
                    ]
                }
            }
        }
    ]
```

```
}
```

Pour les autres ressources, telles que les tâches de formation et les tâches de traitement, vous pouvez rendre les balises obligatoires en appliquant la politique suivante :

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "EnforceTagsForJobs",
            "Effect": "Allow",
            "Action": [
                "sagemaker:CreateTrainingJob",
                "sagemaker:CreateProcessingJob",
            ],
            "Resource": "*",
            "Condition": {
                "ForAnyValue:StringEquals": {
                    "aws:TagKeys": [
                        "studiouserid"
                    ]
                }
            }
        }
    ]
}
```

## Accès root dans SageMaker AI Studio

Dans SageMaker AI Studio, le bloc-notes s'exécute dans un conteneur Docker qui, par défaut, n'a pas d'accès root à l'instance hôte. De même, à l'exception de l'exécution en tant qu'utilisateur par défaut, toutes les autres plages d'identifiants utilisateur à l'intérieur du conteneur sont mappées en tant qu'utilisateur non privilégié IDs sur l'instance hôte elle-même. Par conséquent, la menace d'augmentation des privilèges est limitée au conteneur de blocs-notes lui-même.

Lorsque vous créez des images personnalisées, vous souhaiterez peut-être accorder à votre utilisateur des autorisations non root pour des contrôles plus stricts, par exemple en évitant d'exécuter des processus indésirables en tant que root ou en installant des packages accessibles au public. Dans ce cas, vous pouvez créer l'image à exécuter en tant qu'utilisateur non root dans le Dockerfile. Que vous créiez l'utilisateur en tant que root ou non root, vous devez vous assurer que c'est le UID/GID of the user is identical to the UID/GID cas AppImageConfigpour l'application

personnalisée, qui crée la configuration permettant à l' SageMaker IA d'exécuter une application à l'aide de l'image personnalisée. Par exemple, si votre Dockerfile est conçu pour un utilisateur non root, tel que celui-ci :

```
ARG NB_UID="1000"
ARG NB_GID="100"
...
USER $NB_UID
```

Le `AppImageConfig` fichier doit le mentionner UID et GID dans son dossier `KernelGatewayConfig` :

```
{
    "KernelGatewayImageConfig": {
        "FileSystemConfig": {
            "DefaultUid": 1000,
            "DefaultGid": 100
        }
    }
}
```

Les GID valeursUID/acceptables pour les images personnalisées sont 0/0 et 1000/100 pour les images Studio. Pour des exemples de création d'images personnalisées et des `AppImageConfig` paramètres associés, consultez ce référentiel Github.

Pour éviter que les utilisateurs n'altèrent cela, n'accordez pas de `CreateAppImageConfig` `DeleteAppImageConfig` droits ou d'autorisations aux utilisateurs de blocs-notes SageMaker AI Studio. `UpdateAppImageConfig`

# Gestion du réseau

Pour configurer le domaine SageMaker AI Studio, vous devez spécifier le VPC réseau, les sous-réseaux et les groupes de sécurité. Lorsque vous spécifiez les sous-réseaux VPC et, assurez-vous d'allouer en IPs tenant compte du volume d'utilisation et de la croissance attendue décrits dans les sections suivantes.

## VPCplanification du réseau

VPCLes sous-réseaux clients associés au domaine SageMaker AI Studio doivent être créés avec la plage de routage inter-domaines sans classe (CIDR) appropriée, en fonction des facteurs suivants :

- Nombre d'utilisateurs.
- Nombre d'applications par utilisateur.
- Nombre de types d'instances uniques par utilisateur.
- Nombre moyen d'instances de formation par utilisateur.
- Pourcentage de croissance attendu.

SageMaker L'IA et les AWS services participants injectent des interfaces réseau élastiques (ENI) dans le VPC sous-réseau du client pour les cas d'utilisation suivants :

- Amazon EFS injecte un ENI for et une cible de EFS montage pour le domaine SageMaker AI (une adresse IP par sous-net/zone de disponibilité attachée au domaine SageMaker AI).
- SageMaker AI Studio injecte un ENI pour chaque instance unique utilisée par un profil utilisateur ou un espace partagé. Par exemple :
  - Si un profil utilisateur exécute une application serveur Jupyter par défaut (une instance « système »), une application Data Science et une application Base Python (toutes deux exécutées sur une `ml.t3.medium` instance), Studio injecte deux adresses IP.
  - Si un profil utilisateur exécute une application de serveur Jupyter par défaut (une instance « système »), une application Tensorflow (sur une `ml.g4dn.xlarge` instance) et une GPU application Data Wrangler (sur une `ml.m5.4xlarge` instance), Studio injecte trois adresses IP.
- Un ENI pour chaque VPC point de terminaison dans les VPC sous-réseaux de domaine/zones de disponibilité est injecté (quatre IPs pour les points de VPC terminaison SageMaker AI ; environ six IPs pour les points de VPC terminaison des services participants tels que S3, et.) ECR CloudWatch

- Si les tâches de formation et de traitement de l' SageMaker IA sont lancées avec la même VPC configuration, chaque tâche nécessite deux adresses IP par instance.

> **Note**
>
> VPCles paramètres d' SageMaker AI Studio, tels que les sous-réseaux et le trafic VPC réservé, ne sont pas automatiquement transmis aux tâches de formation/de traitement créées à partir d'AI Studio. SageMaker L'utilisateur doit configurer les VPC paramètres et l'isolation du réseau selon les besoins lorsqu'il appelle le Create*JobAPIs. Reportez-vous à la section Exécuter des conteneurs d'entraînement et d'inférence en mode sans Internet pour plus d'informations.

Scénario : un data scientist réalise des expériences sur deux types d'instances différents

Dans ce scénario, supposons qu'un domaine SageMaker AI soit configuré en mode trafic VPC uniquement. Certains VPC points de terminaison sont configurés, tels que SageMaker AIAPI, SageMaker AI Runtime, Amazon S3 et AmazonECR.

Un data scientist réalise des expériences sur des blocs-notes Studio, s'exécute sur deux types d'instances différents (par exemple, `ml.t3.medium` et`ml.m5.large`) et lance deux applications dans chaque type d'instance.

Supposons que le data scientist exécute également simultanément une tâche de formation avec la même VPC configuration sur une `ml.m5.4xlarge` instance.

Dans ce scénario, le service SageMaker AI Studio injectera ENIs comme suit :

Tableau 1 — ENIs Injecté au client VPC pour un scénario d'expérimentation

| Entité | Cible | ENIinjecté | Remarques | Niveau |
|---|---|---|---|---|
| EFScible de montage | VPCsous-réseaux | Trois | Trois AZs / subnets | Domaine |
| Points de terminaison VPC | VPCsous-réseaux | 30 | Trois AZs / subnets de 10 chacun VPCE | Domaine |

| Entité | Cible | ENIinjecté | Remarques | Niveau |
| --- | --- | --- | --- | --- |
| Serveur Jupyter | Sous-réseau VPC | Un | Une adresse IP par instance | Utilisateur |
| KernelGateway appli | Sous-réseau VPC | Deux | Une adresse IP par type d'instance | Utilisateur |
| Entraînement | Sous-réseau VPC | Deux | Deux IPs par instance de formation<br><br>Cinq IPs par instance de formation si elle EFAest utilisée | Utilisateur |

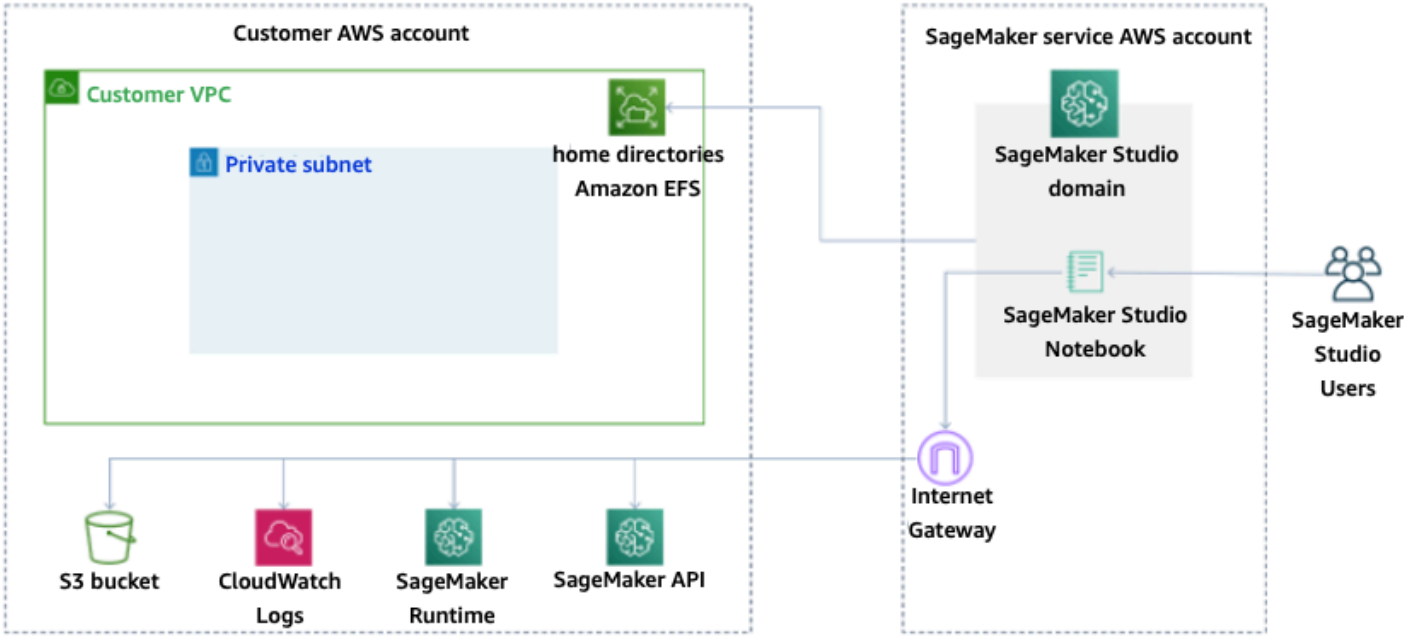
Dans ce scénario, 38 personnes sont IPs consommées au total par le client, VPC dont 33 IPs sont partagées entre les utilisateurs au niveau du domaine et cinq IPs sont consommées au niveau de l'utilisateur. Si 100 utilisateurs ayant des profils utilisateur similaires dans ce domaine effectuent ces activités simultanément, vous consommerez cinq x 100 = 500 IPs au niveau utilisateur, en plus de la consommation IP au niveau du domaine, qui est de 11 IPs par sous-réseau, pour un total de 511. IPs Pour ce scénario, vous devez créer le VPC sous-réseau CIDR avec /22 qui allouera 1024 adresses IP, avec de la marge de croissance.

## VPCoptions réseau

Un domaine SageMaker AI Studio prend en charge la configuration du VPC réseau avec l'une des options suivantes :

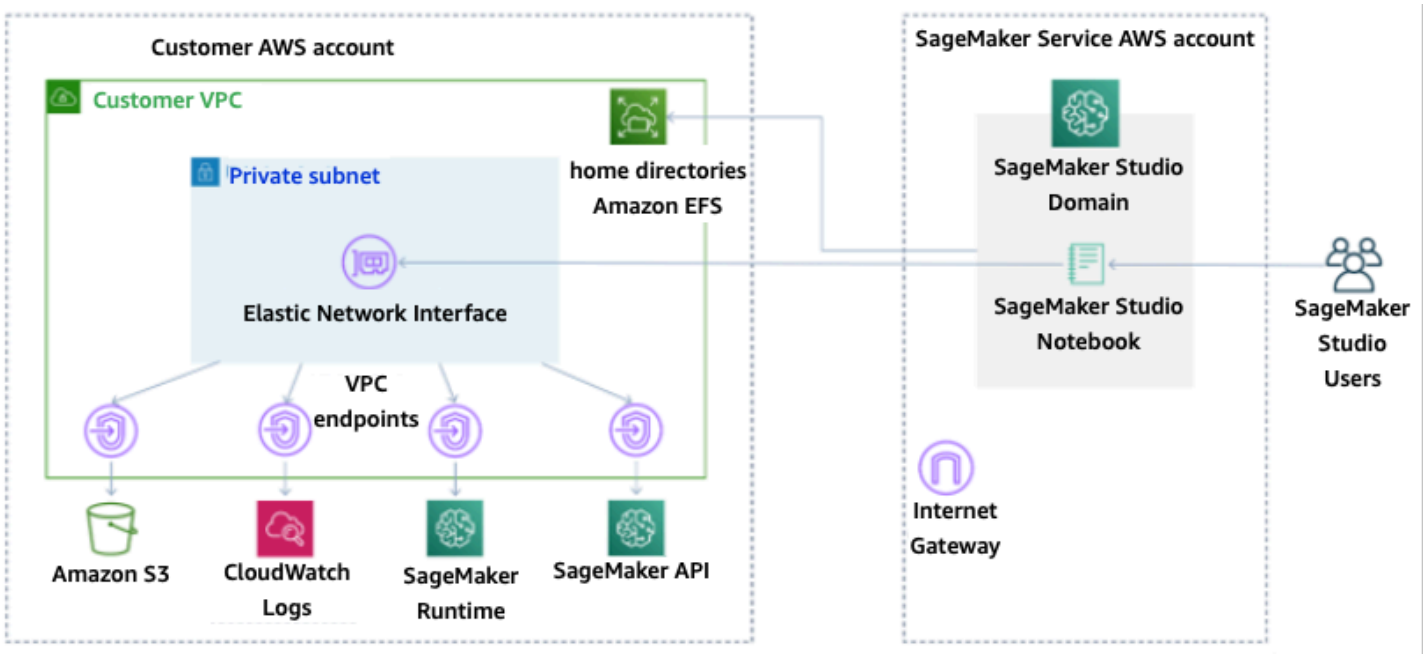
- Internet public uniquement
- VPC uniquement

L'option Internet public uniquement permet aux API services d' SageMaker IA d'utiliser l'Internet public via la passerelle Internet fournie dans le compte de service d' SageMaker IAVPC, gérée par le compte de service AI, comme le montre le schéma suivant :




Mode par défaut : accès à Internet via un compte de service SageMaker AI

La VPCseule option désactive le routage Internet à partir du compte de service VPC géré par l' SageMaker IA et permet au client de configurer le trafic à acheminer via des VPC points de terminaison, comme le montre le schéma suivant :




VPCmode uniquement : pas d'accès à Internet via un compte de service SageMaker AI

Pour un domaine configuré en mode VPC uniquement, configurez un groupe de sécurité par profil utilisateur afin de garantir une isolation complète des instances sous-jacentes. Chaque domaine d'un AWS compte peut avoir sa propre VPC configuration et son propre mode Internet. Pour plus de détails concernant la configuration du VPC réseau, reportez-vous à Connect SageMaker AI Studio Notebooks in a VPC to External Resources.

# Limites

- Après la création d'un domaine SageMaker AI Studio, vous ne pouvez pas associer de nouveaux sous-réseaux au domaine.
- Le type de VPC réseau (Internet public uniquement ou VPCuniquement) ne peut pas être modifié.

# Protection des données

Avant de concevoir l'architecture d'une charge de travail ML, les pratiques fondamentales qui influencent la sécurité doivent être mises en place. Par exemple, la classification des données permet de classer les données en fonction de leur niveau de sensibilité, et le chiffrement protège les données en les rendant incompréhensibles pour tout accès non autorisé. Ces méthodes sont importantes, car elles répondent à des objectifs tels que la prévention des erreurs de manipulation ou le respect des obligations réglementaires.

SageMaker AI Studio propose plusieurs fonctionnalités pour protéger les données au repos et en transit. Cependant, comme décrit dans le modèle de responsabilitéAWS partagée, les clients sont tenus de garder le contrôle sur le contenu hébergé sur l'infrastructure AWS mondiale. Dans cette section, nous décrivons comment les clients peuvent utiliser ces fonctionnalités pour protéger leurs données.

## Protégez les données au repos

Pour protéger vos blocs-notes SageMaker AI Studio ainsi que vos données de création de modèles et vos artefacts, l' SageMaker IA chiffre les blocs-notes, ainsi que les résultats des tâches de formation et de transformation par lots. SageMaker L'IA les chiffre par défaut à l'aide de la cléAWS gérée pour Amazon S3. Cette clé AWS gérée pour Amazon S3 ne peut pas être partagée pour un accès entre comptes. Pour l'accès entre comptes, spécifiez votre clé gérée par le client lors de la création de ressources d' SageMaker IA afin qu'elle puisse être partagée pour un accès entre comptes.

Avec SageMaker AI Studio, les données peuvent être stockées aux emplacements suivants :

- Compartiment S3 : lorsqu'un bloc-notes partageable est activé, SageMaker AI Studio partage les instantanés et les métadonnées du bloc-notes dans un compartiment S3.
- EFSvolume — SageMaker AI Studio attache un EFS volume à votre domaine pour stocker des blocs-notes et des fichiers de données. Ce EFS volume persiste même après la suppression du domaine.
- EBSvolume : EBS est attaché à l'instance sur laquelle le bloc-notes s'exécute. Ce volume est conservé pendant toute la durée de l'instance.

## Chiffrement au repos avec AWS KMS

- Vous pouvez transmettre votre AWS KMS clé pour chiffrer un EBS volume attaché à des ordinateurs portables, à des formations, à des réglages, à des tâches de transformation par lots et à des terminaux.
- Si vous ne spécifiez pas de KMS clé, SageMaker AI chiffre à la fois les volumes du système d'exploitation (OS) et les volumes de données ML à l'aide d'une clé gérée par le systèmeKMS.
- Les données sensibles qui doivent être chiffrées avec une KMS clé pour des raisons de conformité doivent être stockées dans le volume de stockage ML ou dans Amazon S3, les deux pouvant être chiffrés à l'aide d'une KMS clé que vous spécifiez.

# Protéger les données en transit

SageMaker AI Studio garantit que les artefacts du modèle ML et les autres artefacts du système sont chiffrés en transit et au repos. Les demandes adressées à l' SageMaker IA API et à la console sont effectuées via une connexion sécurisée (SSL). Certaines données intra-réseau en transit (au sein de la plateforme de service) ne sont pas chiffrées. Cela consiste notamment à :

- Communications de commande et de contrôle entre le plan de contrôle de service et les instances de tâche d'entraînement (pas les données client).
- Communications entre les nœuds dans le cadre de tâches de traitement et de formation distribuées (intra-réseau).

Vous pouvez toutefois choisir de chiffrer les communications entre les nœuds d'un cluster d'entraînement. L'activation du chiffrement du trafic entre conteneurs peut augmenter la durée de l'entraînement, surtout si vous utilisez des algorithmes de deep learning distribués.

Par défaut, Amazon SageMaker AI exécute des tâches de formation dans un Amazon VPC pour garantir la sécurité de vos données. Vous pouvez ajouter un niveau de sécurité supplémentaire pour protéger vos conteneurs de formation et vos données en configurant un système privéVPC. En outre, vous pouvez configurer votre domaine SageMaker AI Studio pour qu'il s'exécute en mode VPC uniquement et configurer des VPC points de terminaison pour acheminer le trafic sur un réseau privé sans le faire sortir par Internet.

# Garde-corps de protection des données

## Chiffrez les volumes d'hébergement SageMaker AI au repos

Utilisez la politique suivante pour appliquer le chiffrement lors de l'hébergement d'un point de terminaison SageMaker AI pour l'inférence en ligne :

```
{
   "Version": "2012-10-17",
   "Statement": [
     {
         "Sid": "Encryption",
         "Effect": "Allow",
         "Action": [
             "sagemaker:CreateEndpointConfig"
         ],
         "Resource": "*",
         "Condition": {
             "Null": {
                 "sagemaker:VolumeKmsKey": "false"
             }
         }
     }
   ]
 }
```

## Chiffrer les compartiments S3 utilisés lors de la surveillance des modèles

Model Monitoring capture les données envoyées à votre point de terminaison SageMaker AI et les stocke dans un compartiment S3. Lorsque vous configurez la configuration de capture de données, vous devez chiffrer le compartiment S3. Il n'existe actuellement aucun contrôle compensatoire pour cela.

Outre la capture des résultats des terminaux, le service Model Monitoring vérifie la dérive par rapport à une base de référence prédéfinie. Vous devez chiffrer les sorties et les volumes de stockage intermédiaires utilisés pour surveiller la dérive.

```
{
   "Version": "2012-10-17",
   "Statement": [
```

```
    {
        "Sid": "Encryption",
        "Effect": "Allow",
        "Action": [
            "sagemaker:CreateMonitoringSchedule",
            "sagemaker:UpdateMonitoringSchedule"
        ],
        "Resource": "*",
        "Condition": {
            "Null": {
                "sagemaker:VolumeKmsKey": "false",
                "sagemaker:OutputKmsKey": "false"
            }
        }
    }
  ]
}
```

## Chiffrer un volume de stockage de domaine SageMaker AI Studio

Appliquez le chiffrement au volume de stockage attaché au domaine Studio. Cette politique oblige l'utilisateur à fournir un CMK pour chiffrer les volumes de stockage attachés aux domaines du studio.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "EncryptDomainStorage",
            "Effect": "Allow",
            "Action": [
                "sagemaker:CreateDomain"
            ],
            "Resource": "*",
            "Condition": {
                "Null": {
                    "sagemaker:VolumeKmsKey": "false"
                }
            }
        }
    ]
}
```

## Chiffrez les données stockées dans S3 qui sont utilisées pour partager des blocs-notes

Voici la politique qui permet de chiffrer toutes les données stockées dans le compartiment utilisé pour partager des blocs-notes entre les utilisateurs d'un domaine SageMaker AI Studio :

```
{
     "Version": "2012-10-17",
     "Statement": [
         {
             "Sid": "EncryptDomainSharingS3Bucket",
             "Effect": "Allow",
             "Action": [
                 "sagemaker:CreateDomain",
                 "sagemaker:UpdateDomain"
             ],
             "Resource": "*",
             "Condition": {
                 "Null": {
                     "sagemaker:DomainSharingOutputKmsKey": "false"
                 }
             }
         }
     ]
 }
```

# Limites

- Une fois qu'un domaine est créé, vous ne pouvez pas mettre à jour le EFS volume de stockage attaché avec une AWS KMS clé personnalisée.
- Vous ne pouvez pas mettre à jour les tâches de formation/de traitement ou les configurations des terminaux à l'aide de KMS clés une fois qu'elles ont été créées.

# Journalisation et surveillance

Pour vous aider à déboguer vos tâches de compilation, vos tâches de traitement, vos tâches de formation, vos points de terminaison, vos tâches de transformation, vos instances de bloc-notes et vos configurations du cycle de vie des instances de bloc-notes, tout ce qu'un conteneur d'algorithmes, un conteneur de modèles ou une configuration du cycle de vie d'une instance de bloc-notes envoie à stdout ou stderr est également envoyé à Amazon Logs. CloudWatch Vous pouvez surveiller SageMaker AI Studio à l'aide d'Amazon CloudWatch, qui collecte les données brutes et les traite en indicateurs lisibles en temps quasi réel. Ces statistiques sont conservées pendant 15 mois, afin que vous puissiez accéder aux informations historiques et avoir une meilleure idée des performances de votre application ou service Web.

## Se connecter avec CloudWatch

Le processus de science des données étant intrinsèquement expérimental et itératif, il est essentiel de consigner les activités telles que l'utilisation des ordinateurs portables, le temps d'exécution des tâches de formation/de traitement, les indicateurs de formation et les indicateurs de service aux terminaux tels que la latence d'invocation. Par défaut, SageMaker AI publie des métriques dans les CloudWatch journaux, et ces journaux peuvent être chiffrés à l'aide de clés gérées par le client à l'aide de. AWS KMS

Vous pouvez également utiliser des VPC terminaux pour envoyer des journaux CloudWatch sans utiliser l'Internet public. Vous pouvez également définir des alarmes qui surveillent certains seuils et envoient des notifications ou prennent des mesures lorsque ces seuils sont atteints. Pour plus d'informations, consultez le guide de CloudWatch l'utilisateur Amazon.

SageMaker AI crée un groupe de journaux unique pour Studio, sous `/aws/sagemaker/studio`. Chaque profil utilisateur et chaque application ont leur propre flux de journal dans ce groupe de journaux, et les scripts de configuration du cycle de vie ont également leur propre flux de journal. Par exemple, un profil utilisateur nommé « studio-user » associé à une application Jupyter Server associée à un script de cycle de vie, et à une application Data Science Kernel Gateway contient les flux de journaux suivants :

```
/aws/sagemaker/studio/<domain-id>/studio-user/JupyterServer/default
```

```
/aws/sagemaker/studio/<domain-id>/studio-user/JupyterServer/default/
LifecycleConfigOnStart
```

```
/aws/sagemaker/studio/<domain-id>/studio-user/KernelGateway/datascience-app
```

Pour que l' SageMaker IA puisse envoyer CloudWatch des journaux en votre nom, l'appelant Training/Processing/Transform doit APIs disposer des autorisations suivantes :

```
{
     "Version": "2012-10-17",
     "Statement": [
         {
             "Action": [
                 "logs:CreateLogDelivery",
                 "logs:CreateLogGroup",
                 "logs:CreateLogStream",
                 "logs:DeleteLogDelivery",
                 "logs:Describe*",
                 "logs:GetLogEvents",
                 "logs:GetLogDelivery",
                 "logs:ListLogDeliveries",
                 "logs:PutLogEvents",
                 "logs:PutResourcePolicy",
                 "logs:UpdateLogDelivery"
             ],
             "Resource": "*",
             "Effect": "Allow"
         }
     ]
 }
```

Pour chiffrer ces journaux avec une AWS KMS clé personnalisée, vous devez d'abord modifier la politique de clé afin de permettre au CloudWatch service de chiffrer et de déchiffrer la clé. Une fois que vous avez créé une AWS KMS clé de chiffrement du journal, modifiez la politique de clé pour inclure les éléments suivants :

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "logs.region.amazonaws.com"
            },
            "Action": [
```

```
                "kms:Encrypt*",
                "kms:Decrypt*",
                "kms:ReEncrypt*",
                "kms:GenerateDataKey*",
                "kms:Describe*"
            ],
            "Resource": "*",
            "Condition": {
                "ArnLike": {
                    "kms:EncryptionContext:aws:logs:arn": "arn:aws:logs:region:account-
id:*"
                }
            }
        }
    ]
}
```

Notez que vous pouvez toujours utiliser `ArnEquals` et fournir un nom de ressource Amazon spécifique (ARN) pour le CloudWatch journal que vous souhaitez chiffrer. Nous montrons ici que vous pouvez utiliser cette clé pour chiffrer tous les journaux d'un compte pour plus de simplicité. En outre, les points de terminaison de formation, de traitement et de modélisation publient des métriques concernant l'utilisation de l'instance CPU et de la mémoire, la latence d'invocation de l'hébergement, etc. Vous pouvez également configurer Amazon SNS pour informer les administrateurs des événements lorsque certains seuils sont dépassés. Le consommateur participant à la formation et au traitement APIs doit disposer des autorisations suivantes :

```
{
     "Version": "2012-10-17",
     "Statement": [
         {
             "Action": [
                 "cloudwatch:DeleteAlarms",
                 "cloudwatch:DescribeAlarms",
                 "cloudwatch:GetMetricData",
                 "cloudwatch:GetMetricStatistics",
                 "cloudwatch:ListMetrics",
                 "cloudwatch:PutMetricAlarm",
                 "cloudwatch:PutMetricData",
                 "sns:ListTopics"
             ],
             "Resource": "*",
             "Effect": "Allow",
```

```
            "Condition": {
                "StringLike": {
                    "cloudwatch:namespace": "aws/sagemaker/*"
                }
            }

        },
        {
            "Action": [
                "sns:Subscribe",
                "sns:CreateTopic"
            ],
            "Resource": [
                "arn:aws:sns:*:*:*SageMaker*",
                "arn:aws:sns:*:*:*Sagemaker*",
                "arn:aws:sns:*:*:*sagemaker*"
            ],
            "Effect": "Allow"
        }
    ]
}
```

## Audit avec AWS CloudTrail

Pour améliorer votre niveau de conformité, effectuez un audit de toutes vos APIs activités avec AWS CloudTrail. Par défaut, toutes les SageMaker IA APIs sont connectées AWS CloudTrail. Vous n'avez pas besoin d'IAMautorisations supplémentaires pour l'activer CloudTrail.

Toutes les actions de l' SageMaker IA, à l'exception de `InvokeEndpoint` et`InvokeEndpointAsync`, sont enregistrées CloudTrail et documentées dans les opérations. Par exemple, les appels aux `CreateTrainingJob``CreateEndpoint`, et `CreateNotebookInstance` les actions génèrent des entrées dans les fichiers CloudTrail journaux.

Chaque entrée d' CloudTrail événement contient des informations sur l'auteur de la demande. Les informations relatives à l'identité permettent de déterminer les éléments suivants :

- Si la demande a été effectuée avec les informations d'identification utilisateur racine ou AWS IAM.
- Si la demande a été effectuée avec les informations d'identification de sécurité temporaires d'un rôle ou d'un utilisateur fédéré.
- Si la demande a été faite par un autre AWS service. Pour un exemple d'événement, reportez-vous à la CloudTrail documentation Log SageMaker AI API Calls.

Par défaut, CloudTrail enregistre le nom du rôle d'exécution Studio du profil utilisateur comme identifiant pour chaque événement. Cela fonctionne si chaque utilisateur a son propre rôle d'exécution. Si plusieurs utilisateurs partagent le même rôle d'exécution, vous pouvez utiliser la `sourceIdentity` configuration pour propager le nom du profil utilisateur Studio à CloudTrail. Reportez-vous à la section Surveillance de l'accès aux ressources utilisateur depuis Amazon SageMaker AI Studio pour activer `sourceIdentity` cette fonctionnalité. Dans un espace partagé, toutes les actions font référence à l'espace ARN en tant que source, et vous ne pouvez pas effectuer d'audit`sourceIdentity`.

# Attribution des coûts

SageMaker AI Studio intègre des fonctionnalités pour aider les administrateurs à suivre les dépenses de leurs domaines individuels, de leurs espaces partagés et de leurs utilisateurs.

## Marquage automatique

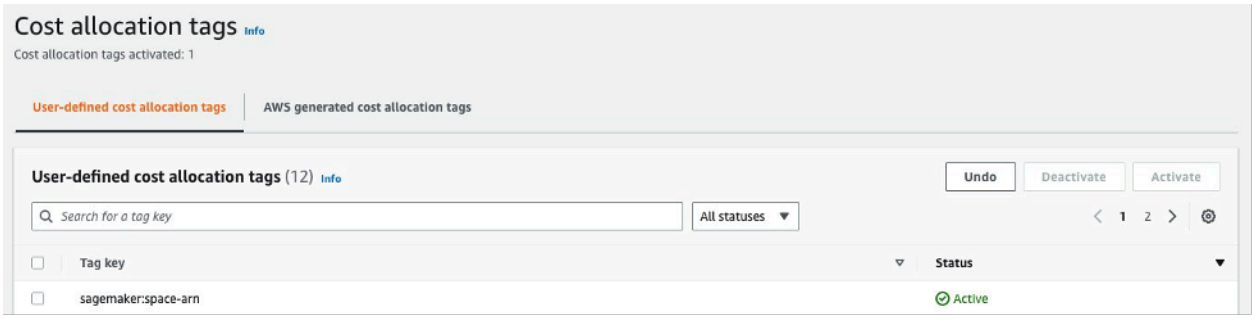
SageMaker AI Studio étiquette désormais automatiquement les nouvelles SageMaker ressources telles que les tâches de formation, les tâches de traitement et les applications du noyau avec leurs balises respectives`sagemaker:domain-arn`. À un niveau plus détaillé, l' SageMaker IA étiquette également la ressource avec le `sagemaker:user-profile-arn` ou `sagemaker:space-arn` pour désigner le créateur principal de la ressource.

SageMaker Les EFS volumes de domaine AI sont étiquetés avec une clé nommée `ManagedByAmazonSageMakerResource` avec la valeur du domaineARN. Ils ne disposent pas de balises granulaires permettant de comprendre l'utilisation de l'espace au niveau de chaque utilisateur. Les administrateurs peuvent toutefois associer le EFS volume à une EC2 instance pour une surveillance personnalisée.

## Suivi des coûts

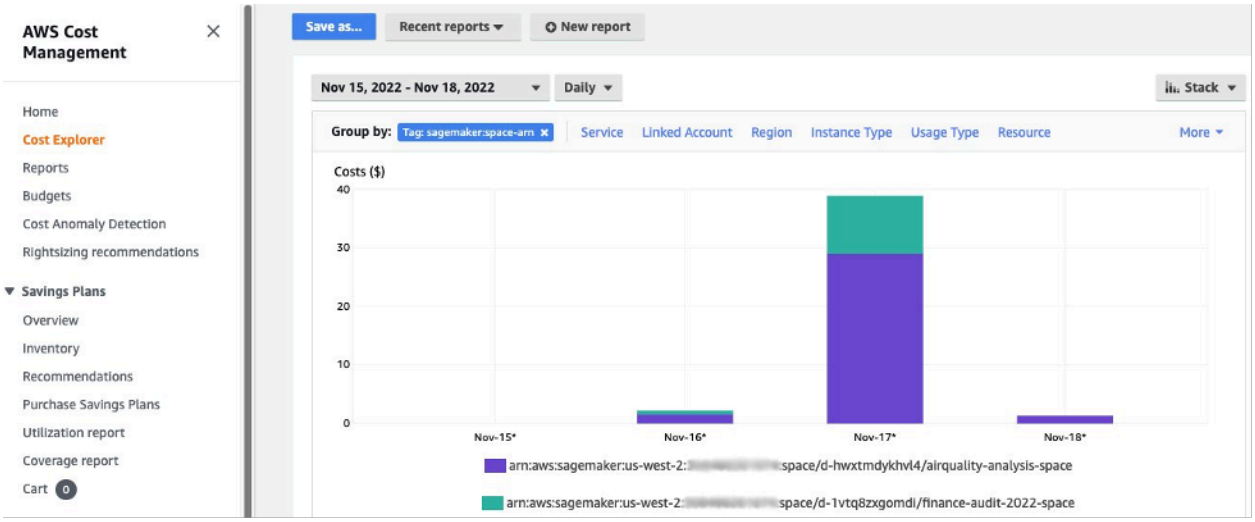
Les balises automatisées permettent aux administrateurs de suivre, de signaler et de surveiller vos dépenses en ML grâce à out-of-the-box des solutions telles que AWS Cost Exploreret AWS Budgets, ainsi qu'à des solutions personnalisées basées sur les données des rapports sur les AWS coûts et l'utilisation (CURs).

Pour utiliser les balises jointes à des fins d'analyse des coûts, elles doivent d'abord être activées dans la section Balises de répartition des coûts de la AWS Billing console. L'affichage des balises dans le panneau des balises de répartition des coûts peut prendre jusqu'à 24 heures. Vous devez donc créer une ressource d' SageMaker IA avant de les activer.



Espace ARN activé en tant que balises de répartition des coûts dans Cost Explorer

Une fois que vous avez activé une balise de répartition des coûts, vous AWS commencerez à suivre vos ressources étiquetées, et après 24 à 48 heures, les balises apparaîtront sous forme de filtres sélectionnables dans l'explorateur de coûts.




Coûts regroupés par espace partagé pour un exemple de domaine

# Contrôle des coûts

Lorsque le premier utilisateur d' SageMaker AI Studio est intégré, SageMaker AI crée un EFS volume pour le domaine. Des frais de stockage sont engagés pour ce EFS volume car les blocs-notes et les fichiers de données sont stockés dans le répertoire personnel de l'utilisateur. Lorsque l'utilisateur lance des blocs-notes Studio, ceux-ci sont lancés pour les instances de calcul qui exécutent les blocs-notes. Reportez-vous à la tarification d'Amazon SageMaker AI pour une ventilation détaillée des coûts.

Les administrateurs peuvent contrôler les coûts de calcul en spécifiant la liste des instances qu'un utilisateur peut créer, en utilisant IAM les politiques mentionnées dans la section Garde-fous courants. En outre, nous recommandons aux clients d'utiliser l'extension d'arrêt automatique SageMaker AI Studio pour réduire les coûts en fermant automatiquement les applications inactives. Cette extension de serveur interroge régulièrement les applications en cours d'exécution par profil utilisateur et arrête les applications inactives en fonction d'un délai défini par l'administrateur.

Pour définir cette extension pour tous les utilisateurs de votre domaine, vous pouvez utiliser une configuration de cycle de vie telle que décrite dans la section Personnalisation. En outre, vous

pouvez également utiliser le vérificateur d'extension pour vous assurer que l'extension est installée sur tous les utilisateurs de votre domaine.

# Personnalisation

## Configuration du cycle de vie

Les configurations du cycle de vie sont des scripts shell initiés par des événements du cycle de vie d' SageMaker AI Studio, tels que le démarrage d'un nouveau bloc-notes SageMaker AI Studio. Vous pouvez utiliser ces scripts shell pour automatiser la personnalisation de vos environnements SageMaker AI Studio, comme l'installation de packages personnalisés, l'extension Jupyter pour l'arrêt automatique des applications de bloc-notes inactives et la configuration de Git. Pour obtenir des instructions détaillées sur la façon de créer des configurations de cycle de vie, consultez ce blog : Personnaliser Amazon SageMaker AI Studio à l'aide des configurations de cycle de vie.

## Images personnalisées pour les ordinateurs portables SageMaker AI Studio

Les blocs-notes Studio sont fournis avec un ensemble d'images prédéfinies, qui comprennent Amazon AI SageMaker Python SDK et la dernière version du IPython moteur d'exécution ou du noyau. Grâce à cette fonctionnalité, vous pouvez ajouter vos propres images personnalisées aux blocs-notes Amazon SageMaker AI. Ces images sont ensuite accessibles à tous les utilisateurs authentifiés dans le domaine.

Les développeurs et les data scientists peuvent avoir besoin d'images personnalisées pour différents cas d'utilisation :

- Accès à des versions spécifiques ou récentes de frameworks ML populaires tels que TensorFlowMXNet, PyTorch, ou autres.
- Intégrez du code personnalisé ou des algorithmes développés localement dans les blocs-notes SageMaker AI Studio pour accélérer les itérations et l'apprentissage des modèles.
- Accès aux lacs de données ou aux magasins de données sur site viaAPIs. Les administrateurs doivent inclure les pilotes correspondants dans l'image.
- Accès à un environnement d'exécution principal (également appelé noyau), autre que IPython (tel que R, Julia ou autres). Vous pouvez également utiliser l'approche décrite pour installer un noyau personnalisé.

Pour obtenir des instructions détaillées sur la création d'une image personnalisée, reportez-vous à la section Création d'une image SageMaker AI personnalisée.

## JupyterLab extensions

Avec SageMaker AI Studio JuypterLab 3 Notebook, vous pouvez tirer parti de la communauté toujours croissante d'extensions open source JupyterLab. Cette section met en évidence quelques-unes qui s'intègrent naturellement dans le flux de travail des développeurs d' SageMaker IA, mais nous vous encourageons à parcourir les extensions disponibles ou même à créer les vôtres.

JupyterLab 3 facilite désormais considérablement le processus d'empaquetage et d'installation des extensions. Vous pouvez installer les extensions susmentionnées par le biais de scripts bash. Par exemple, dans SageMaker AI Studio, ouvrez le terminal système à partir du lanceur Studio et exécutez les commandes suivantes. En outre, vous pouvez automatiser l'installation de ces extensions à l'aide de configurations de cycle de vie afin qu'elles soient conservées entre les redémarrages de Studio. Vous pouvez le configurer pour tous les utilisateurs du domaine ou au niveau d'un utilisateur individuel.

Par exemple, pour installer une extension pour un navigateur de fichiers Amazon S3, exécutez les commandes suivantes dans le terminal système et assurez-vous d'actualiser votre navigateur :

```
conda init
conda activate studio
pip install jupyterlab_s3_browser
jupyter serverextension enable --py jupyterlab_s3_browser
conda deactivate
restart-jupyter-server
```

Pour plus d'informations sur la gestion des extensions, notamment sur la façon d'écrire des configurations de cycle de vie qui fonctionnent à la fois pour les versions 1 et 3 des JupyterLab ordinateurs portables à des fins de rétrocompatibilité, reportez-vous à la section Installation JupyterLab et extensions Jupyter Server.

## Référentiels Git

SageMaker AI Studio est préinstallé avec une extension Jupyter Git permettant aux utilisateurs de saisir un dépôt Git URL personnalisé, de le cloner dans EFS votre répertoire, d'effectuer des

modifications et de consulter l'historique des validations. Les administrateurs peuvent configurer les dépôts git suggérés au niveau du domaine afin qu'ils apparaissent sous forme de listes déroulantes pour les utilisateurs finaux. Reportez-vous à la section Attacher des dépôts Git suggérés à Studio pour up-to-date obtenir des instructions.

Si un dépôt est privé, l'extension demandera à l'utilisateur de saisir ses informations d'identification dans le terminal à l'aide de l'installation git standard. L'utilisateur peut également stocker les informations d'identification SSH dans son EFS répertoire individuel pour en faciliter la gestion.

## Environnement Conda

SageMaker Les blocs-notes AI Studio utilisent Amazon EFS comme couche de stockage persistante. Les data scientists peuvent utiliser le stockage persistant pour créer des environnements conda personnalisés et utiliser ces environnements pour créer des noyaux. Ces noyaux sont soutenus par et sont persistants entre EFS les redémarrages du noyau, de l'application ou de Studio. Studio sélectionne automatiquement tous les environnements valides sous forme de KernelGateway noyaux.

Le processus de création d'un environnement conda est simple pour un data scientist, mais les noyaux mettent environ une minute à être renseignés dans le sélecteur de noyau. Pour créer un environnement, exécutez ce qui suit dans un terminal système :

```
mkdir -p ~/.conda/envs
conda create --yes -p ~/.conda/envs/custom
conda activate ~/.conda/envs/custom
conda install -y ipykernel
conda config --add envs_dirs ~/.conda/envs
```

Pour obtenir des instructions détaillées, reportez-vous à la section Persist Conda environments to the Studio EFS volume de la section Quatre approches pour gérer les packages Python dans les blocs-notes Amazon SageMaker Studio.

# Conclusion

Dans ce livre blanc, nous avons passé en revue plusieurs bonnes pratiques dans des domaines tels que le modèle d'exploitation, la gestion des domaines, la gestion des identités, la gestion des autorisations, la gestion du réseau, la journalisation, la surveillance et la personnalisation afin de permettre aux administrateurs de la plateforme de configurer et de gérer SageMaker AI Studio Platform.

# Annexe

## Comparaison entre plusieurs locataires

Tableau 2 — Comparaison entre plusieurs locataires

| Multi-domaines | Comptes multiples | Contrôle d'accès basé sur les attributs (ABAC) au sein d'un même domaine |
|---|---|---|
| L'isolation des ressource s est réalisée à l'aide de balises. SageMaker AI Studio étiquette automatiquement toutes les ressources avec le domaine ARN et le profil/es pace utilisateur. ARN | Chaque locataire a son propre compte, il y a donc une isolation absolue des ressources. | L'isolation des ressources est réalisée à l'aide de balises. Les utilisateurs doivent gérer le balisage des ressources créées pourABAC. |
| La liste APIs ne peut pas être limitée par des balises. Le filtrage des ressource s par l'interface utilisateur est effectué sur les espaces partagés, mais les API appels List effectués via le Boto3 AWS CLI ou le Boto3 SDK listeront les ressources de la région. | APIsL'isolation des listes est également possible, puisque les locataires se trouvent dans leurs comptes dédiés. | La liste APIs ne peut pas être limitée par des balises. Lister API les appels passés via le Boto3 AWS CLI ou le Boto3 SDK listera les ressources de la Région. |
| SageMaker Les coûts de calcul et de stockage d'AI Studio par locataire peuvent être facilement surveillés en utilisant Domain ARN comme étiquette de répartition des coûts. | SageMaker Les coûts de calcul et de stockage d'AI Studio par locataire sont faciles à surveiller avec un compte dédié. | SageMaker Les coûts de calcul d'AI Studio par locataire doivent être calculés à l'aide de balises personnalisées.<br><br>SageMaker Les coûts de stockage d'AI Studio ne |

| Multi-domaines | Comptes multiples | Contrôle d'accès basé sur les attributs (ABAC) au sein d'un même domaine |
|---|---|---|
| | | peuvent pas être surveillés par domaine puisque tous les locataires partagent le même EFS volume. |
| Les quotas de service sont définis au niveau du compte, de sorte qu'un seul locataire peut toujours utiliser toutes les ressources. | Les quotas de service peuvent être définis au niveau du compte pour chaque locataire. | Les quotas de service sont définis au niveau du compte, de sorte qu'un seul locataire peut toujours utiliser toutes les ressources. |
| La mise à l'échelle vers plusieurs locataires peut être réalisée par le biais de l'infrast ructure sous forme de code (IaC) ou du Service Catalog. | L'extension à plusieurs locataires implique des Organisations et la vente de plusieurs comptes. | Le dimensionnement nécessite un rôle spécifiqu e au locataire pour chaque nouveau locataire, et les profils utilisateur doivent être étiquetés manuellement avec les noms des locataires. |
| La collaboration entre les utilisateurs au sein d'un locataire est possible grâce à des espaces partagés. | La collaboration entre utilisate urs au sein d'un locataire est possible grâce à des espaces partagés. | Tous les locataires auront accès au même espace partagé pour la collaboration. |

# SageMaker Sauvegarde et restauration de domaines AI Studio

En cas de EFS suppression accidentelle ou lorsqu'un domaine doit être recréé en raison de modifications apportées au réseau ou à l'authentification, suivez ces instructions.

# Option 1 : Sauvegarder à partir d'une sauvegarde existante à EFS l'aide d'un EC2

## SageMaker Sauvegarde du domaine Studio

1. Répertoriez les profils utilisateur et les espaces dans SageMaker Studio (CLI, SDK).
2. Mappez les profils/espaces utilisateur à UIDs on. EFS
   a. Pour chaque utilisateur de la liste de users/spaces, describe the user profile/space (CLI, SDK).
   b. Mappez le profil/espace utilisateur à. `HomeEfsFileSystemUid`
   c. Mappez le profil utilisateur selon `UserSettings['ExecutionRole']` si les utilisateurs ont des rôles d'exécution distincts.
   d. Identifiez le rôle d'exécution de Space par défaut.
3. Créez un nouveau domaine et spécifiez le rôle d'exécution par défaut de Space.
4. Créez des profils utilisateur et des espaces.
   - Pour chaque utilisateur de la liste d'utilisateurs, créez un profil utilisateur (CLI, SDK) à l'aide du mappage des rôles d'exécution.
5. Créez un mappage pour le nouveau EFS etUIDs.
   a. Pour chaque utilisateur figurant dans la liste des utilisateurs, décrivez le profil utilisateur (CLI, SDK).
   b. Associer le profil utilisateur à`HomeEfsFileSystemUid`.
6. Vous pouvez éventuellement supprimer toutes les applications, tous les profils utilisateur, tous les espaces, puis supprimer le domaine.

## Sauvegarde EFS

Pour effectuer une sauvegardeEFS, suivez les instructions suivantes :

1. Lancez l'EC2instance et associez les groupes de sécurité entrants/sortants de l'ancien domaine SageMaker Studio à la nouvelle EC2 instance (autorisez le NFS trafic sur TCP le port 2049). Reportez-vous à la section Connect SageMaker Studio Notebooks dans la section « Ressources externes ». VPC
2. Montez le EFS volume SageMaker Studio sur la nouvelle EC2 instance. Reportez-vous à la section Montage de systèmes de EFS fichiers.
3. Copiez les fichiers sur le stockage EBS local : `>sudo cp -rp /efs /studio-backup:`

a. Attachez les nouveaux groupes de sécurité de domaine à l'EC2instance.

b. Montez le nouveau EFS volume sur l'EC2instance.

c. Copiez les fichiers sur le nouveau EFS volume.

d. Pour chaque utilisateur de la collection de l'utilisateur :

   i. Créez le répertoire :`mkdir new_uid`.

   ii. Copiez les fichiers de l'ancien UID répertoire vers UID le nouveau.

   iii. Changer de propriétaire pour tous les fichiers : `chown <new_UID>` pour tous les fichiers.

## Option 2 : sauvegarde à partir de données existantes à EFS l'aide de S3 et de la configuration du cycle de vie

1. Reportez-vous à la section Migrer votre travail vers une instance de SageMaker bloc-notes Amazon avec Amazon Linux 2.
2. Créez un compartiment S3 pour la sauvegarde (par exemple>`studio-backup`.
3. Répertoriez tous les profils utilisateur dotés de rôles d'exécution.
4. Dans le domaine SageMaker Studio actuel, définissez un LCC script par défaut au niveau du domaine.
   - Dans leLCC, copiez tout dans `/home/sagemaker-user` le préfixe du profil utilisateur dans S3 (par exemple,`s3://studio-backup/studio-user1`).
5. Redémarrez toutes les applications Jupyter Server par défaut (pour LCC qu'elles soient exécutées).
6. Supprimez toutes les applications, tous les profils utilisateur et tous les domaines.
7. Créez un nouveau domaine SageMaker Studio.
8. Créez de nouveaux profils utilisateur à partir de la liste des profils utilisateur et des rôles d'exécution.
9. Configurez un LCC au niveau du domaine :
   - Dans leLCC, copiez tout ce qui se trouve dans le préfixe du profil utilisateur dans S3 vers `/home/sagemaker-user`
10. Créez des applications Jupyter Server par défaut pour tous les utilisateurs avec la LCCconfiguration (CLI, SDK).

# SageMaker Accès au studio à l'aide d'une SAML assertion

Configuration de la solution :

1. Créez une SAML application dans votre IdP externe.
2. Configurez l'IdP externe en tant que fournisseur d'identité dans. IAM
3. Créez une fonction `SAMLValidator` Lambda accessible à l'IdP (via une fonction URL ou une passerelle). API
4. Créez une fonction `GeneratePresignedUrl` Lambda et une API passerelle pour accéder à la fonction.
5. Créez un IAM rôle que les utilisateurs peuvent assumer pour appeler la API passerelle. Ce rôle doit être transmis en SAML assertion sous forme d'attribut au format suivant :
   - Nom de l'attribut : https://aws.amazon.com/SAML/ Attributs/Rôle
   - Valeur de l'attribut :`<IdentityProviderARN>`, `<RoleARN>`
6. Mettez à jour le SAML point de terminaison Assertion Consumer Service (ACS) sur l'`SAMLValidator`invocationURL.

SAMLexemple de code de validateur :

```
import requests
import os
import boto3
from urllib.parse import urlparse, parse_qs
import base64
import requests
from aws_requests_auth.aws_auth import AWSRequestsAuth
import json


# Config for calling AssumeRoleWithSAML
idp_arn = "arn:aws:iam::0123456789:saml-provider/MyIdentityProvider"
api_gw_role_arn = 'arn:aws:iam:: 0123456789:role/APIGWAccessRole'
studio_api_url = "abcdef.execute-api.us-east-1.amazonaws.com"
studio_api_gw_path = "https://" + studio_api_url + "/Prod "

# Every customer will need to get SAML Response from the POST call
def get_saml_response(event):
    saml_response_uri = base64.b64decode(event['body']).decode('ascii')
```

```
    request_body = parse_qs(saml_response_uri)
    print(f"b64 saml response: {request_body['SAMLResponse'][0]}")
    return request_body['SAMLResponse'][0]


def lambda_handler(event, context):
    sts = boto3.client('sts')

    # get temporary credentials
    response = sts.assume_role_with_saml(
                    RoleArn=api_gw_role_arn,
                    PrincipalArn=durga_idp_arn,
                    SAMLAssertion=get_saml_response(event)
                )
    auth = AWSRequestsAuth(aws_access_key=response['Credentials']['AccessKeyId'],
                      aws_secret_access_key=response['Credentials']['SecretAccessKey'],
                      aws_host=studio_api_url,
                      aws_region='us-west-2',
                      aws_service='execute-api',
                      aws_token=response['Credentials']['SessionToken'])

    presigned_response = requests.post(
        studio_api_gw_path,
        data=saml_response_data,
        auth=auth)

    return presigned_response
```

# Suggestions de lecture

- Configuration d'environnements d'apprentissage automatique sécurisés et bien gérés sur AWS (AWS blog)
- Configuration d'Amazon SageMaker AI Studio pour les équipes et les groupes avec une isolation complète des ressources (AWS blog)
- Intégration d'Amazon SageMaker AI Studio AWS SSO et d'Okta Universal Directory (blog)AWS
- Comment configurer la SAML version 2.0 pour la fédération de AWS comptes (documentation Okta)
- Création d'une plateforme de Machine Learning d'entreprise sécurisée sur AWS (guide AWS technique)
- Personnalisez Amazon SageMaker AI Studio à l'aide des configurations du cycle de vie (AWS blog)
- Intégrer votre propre image de conteneur personnalisée aux blocs-notes Amazon SageMaker AI Studio (AWS blog)
- Créez des modèles de projets d' SageMaker IA personnalisés — Meilleures pratiques (AWS blog)
- Déploiement de modèles multi-comptes avec Amazon SageMaker AI Pipelines (AWS blog)
- Partie 1 : Comment le NatWest groupe a créé une MLOps plateforme évolutive, sécurisée et durable (AWS blog)
- Secure Amazon SageMaker AI Studio présigné, URLs partie 1 : infrastructure de base (blog)AWS

# Collaborateurs

Les personnes qui ont contribué à ce document incluent :

- Ram Vittal, architecte de solutions ML, Amazon Web Services
- Sean Morgan, architecte de solutions ML, Amazon Web Services
- Durga Sury, architecte de solutions ML, Amazon Web Services

Nous remercions tout particulièrement les personnes suivantes qui ont apporté des idées, des révisions et des points de vue :

- Alessandro Cerè, architecte de solutions d'intelligence artificielle et d'apprentissage automatique, Amazon Web Services
- Sumit Thakur, responsable des produits d' SageMaker intelligence artificielle, Amazon Web Services
- Han Zhang, ingénieur principal en développement logiciel, Amazon Web Services
- Bhadrinath Pani, ingénieur en développement logiciel, Amazon Web Services, Amazon Web Services

# Révisions du document

Pour être informé des mises à jour de ce livre blanc, abonnez-vous au flux RSS.

| Modification | Description | Date |
| --- | --- | --- |
| Livre blanc mis à jour | Les liens rompus ont été corrigés et de nombreuses modifications éditoriales ont été apportées. | 25 avril 2023 |
| Publication initiale | Livre blanc publié. | 19 octobre 2022 |

# Avis

Les clients sont tenus de procéder à leur propre évaluation indépendante des informations contenues dans ce document. Ce document : (a) est fourni à titre informatif uniquement, (b) représente les offres de AWS produits et les pratiques actuelles, qui sont susceptibles d'être modifiées sans préavis, et (c) ne crée aucun engagement ni aucune assurance de la part de AWS ses filiales, fournisseurs ou concédants de licence. AWSles produits ou services sont fournis « tels quels » sans garantie, représentation ou condition d'aucune sorte, expresse ou implicite. Les responsabilités et obligations AWS de ses clients sont régies par AWS des accords, et ce document ne fait partie d'aucun accord conclu entre AWS et ses clients et ne les modifie pas.

© 2022 Amazon Web Services, Inc. ou ses filiales. Tous droits réservés.

# Glossaire AWS

Pour connaître la terminologie la plus récente d’AWS, consultez le Glossaire AWS dans la Référence Glossaire AWS.

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.