AWS Whitepaper

SageMaker Bewährte Methoden für die Studio-Administration

SageMaker Bewährte Methoden für die Studio-Administration: AWS Whitepaper

Table of Contents

SageMaker Bewährte Methoden für die Studio-Administration

Datum der Veröffentlichung: 25. April 2023 (Dokumentversionen)

Überblick

Amazon SageMaker AI Studio bietet eine einzige, webbasierte visuelle Oberfläche, über die Sie alle Entwicklungsschritte des maschinellen Lernens (ML) durchführen können, was die Produktivität von Data-Science-Teams verbessert. SageMaker AI Studio bietet Ihnen vollständigen Zugriff, Kontrolle und Transparenz für jeden Schritt, der zum Erstellen, Trainieren und Evaluieren von Modellen erforderlich ist.

In diesem Whitepaper besprechen wir bewährte Methoden für Themen wie Betriebsmodell, Domänenmanagement, Identitätsmanagement, Berechtigungsmanagement, Netzwerkmanagement, Protokollierung, Überwachung und Anpassung. Die hier erörterten Best Practices sind für die Bereitstellung von SageMaker KI Studio in Unternehmen, einschließlich Bereitstellungen mit mehreren Mandanten, vorgesehen. Dieses Dokument richtet sich an ML-Plattformadministratoren, ML-Ingenieure und ML-Architekten.

Sind Sie Well-Architected?

Das AWS Well-Architected Framework hilft Ihnen dabei, die Vor- und Nachteile der Entscheidungen zu verstehen, die Sie beim Aufbau von Systemen in der Cloud treffen. Die sechs Säulen des Frameworks ermöglichen es Ihnen, bewährte Architekturpraktiken für den Entwurf und Betrieb zuverlässiger, sicherer, effizienter, kostengünstiger und nachhaltiger Systeme kennenzulernen. Mithilfe des AWS Well-Architected Tool, das kostenlos im verfügbar ist AWS Management Console, können Sie Ihre Workloads anhand dieser bewährten Methoden überprüfen, indem Sie für jede Säule eine Reihe von Fragen beantworten.

In der Machine Learning Lens konzentrieren wir uns darauf, wie Sie Ihre Workloads für maschinelles Lernen in der AWS Cloud entwerfen, bereitstellen und gestalten können. Diese Linse ergänzt die im Well-Architected Framework beschriebenen Best Practices.

Einführung

Wenn Sie SageMaker AI Studio als Ihre ML-Plattform verwalten, benötigen Sie Anleitungen zu bewährten Methoden, um fundierte Entscheidungen treffen zu können, damit Sie Ihre ML-Plattform skalieren können, wenn Ihre Workloads wachsen. Beachten Sie bei der Bereitstellung, Operationalisierung und Skalierung Ihrer ML-Plattform Folgendes:

- Wählen Sie das richtige Betriebsmodell und organisieren Sie Ihre ML-Umgebungen so, dass sie Ihre Geschäftsziele erreichen.
- Wählen Sie aus, wie die SageMaker AI Studio-Domänenauthentifizierung für Benutzeridentitäten eingerichtet werden soll, und berücksichtigen Sie dabei die Einschränkungen auf Domänenebene.
- Entscheiden Sie, wie Sie die Identität und Autorisierung Ihrer Benutzer mit der ML-Plattform verbinden möchten, um detaillierte Zugriffskontrollen und Prüfungen zu ermöglichen.
- Erwägen Sie, Berechtigungen und Leitplanken für verschiedene Rollen Ihrer ML-Personas einzurichten.
- Planen Sie Ihre Virtual Private Cloud (VPC) -Netzwerktopologie unter Berücksichtigung der Sensitivität Ihres ML-Workloads, der Anzahl der Benutzer, der Instanztypen, der Apps und der gestarteten Jobs.
- Klassifizieren und schützen Sie Ihre Daten im Ruhezustand und bei der Übertragung mit Verschlüsselung.
- Überlegen Sie, wie Sie verschiedene Anwendungsprogrammierschnittstellen (APIs) und Benutzeraktivitäten protokollieren und überwachen können, um die Einhaltung der Vorschriften zu gewährleisten.
- Passen Sie das SageMaker AI Studio-Notebook-Erlebnis mit Ihren eigenen Bildern und Lebenszyklus-Konfigurationsskripten an.

Betriebsmodell

Ein Betriebsmodell ist ein Framework, das Menschen, Prozesse und Technologien zusammenbringt, um ein Unternehmen dabei zu unterstützen, Geschäftswert auf skalierbare, konsistente und effiziente Weise zu erzielen. Das ML-Betriebsmodell bietet einen standardisierten Produktentwicklungsprozess für Teams im gesamten Unternehmen. Je nach Größe, Komplexität und Geschäftsfaktoren gibt es drei Modelle für die Implementierung des Betriebsmodells:

- Zentralisiertes Data-Science-Team — In diesem Modell sind alle datenwissenschaftlichen Aktivitäten innerhalb eines einzigen Teams oder einer Organisation zentralisiert. Dies ähnelt dem Modell des Center of Excellence (COE), bei dem alle Geschäftsbereiche für datenwissenschaftliche Projekte an dieses Team weitergeleitet werden.
- Dezentrale Data-Science-Teams — In diesem Modell sind die datenwissenschaftlichen Aktivitäten auf verschiedene Geschäftsfunktionen oder -abteilungen verteilt oder basieren auf unterschiedlichen Produktlinien.
- Föderierte Data-Science-Teams — In diesem Modell werden Shared-Services-Funktionen wie Code-Repositorys, CI/CD-Pipelines (Continuous Integration and Continuous Delivery) usw. vom zentralen Team verwaltet, und jede Funktion auf Geschäftseinheit oder Produktebene wird von dezentralen Teams verwaltet. Dies ähnelt dem Hub-and-Spoke-Modell, bei dem jede Geschäftseinheit ihre eigenen Data-Science-Teams hat. Diese Teams der Geschäftsbereiche koordinieren ihre Aktivitäten jedoch mit dem zentralisierten Team.

Bevor Sie sich entscheiden, Ihre erste Studio-Domain für Anwendungsfälle in der Produktion zu starten, sollten Sie Ihr Betriebsmodell und AWS bewährte Methoden für die Organisation Ihrer Umgebung berücksichtigen. Weitere Informationen finden Sie unter Organizing Your AWS Environment Using Multiple Accounts.

Der nächste Abschnitt enthält Anleitungen zur Organisation Ihrer Kontostruktur für die einzelnen Betriebsmodelle.

Empfohlene Kontostruktur

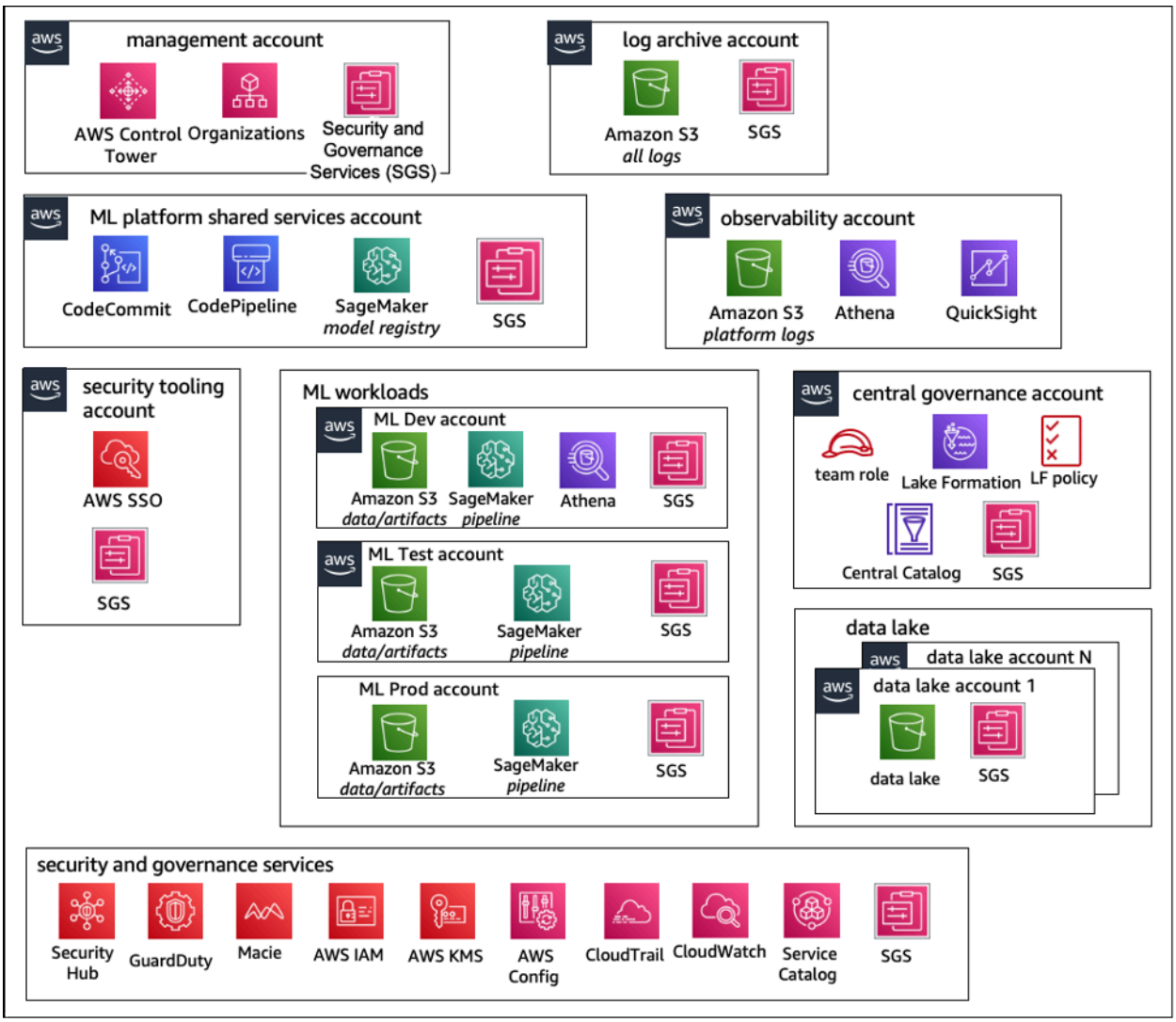
In diesem Abschnitt stellen wir kurz eine Kontostruktur nach dem Betriebsmodell vor, mit der Sie beginnen und die Sie entsprechend den betrieblichen Anforderungen Ihres Unternehmens ändern können. Unabhängig davon, für welches Betriebsmodell Sie sich entscheiden, empfehlen wir die Implementierung der folgenden gängigen bewährten Methoden:

- Verwenden Sie es AWS Control Towerfür die Einrichtung, Verwaltung und Verwaltung Ihrer Konten.
- Zentralisieren Sie Ihre Identitäten mit Ihrem Identity Provider (IdP) und AWS IAMIdentity Center mit einem delegierten Securitiy Tooling-Administratorkonto und ermöglichen Sie den sicheren Zugriff auf Workloads.
- Führen Sie ML-Workloads mit Isolierung auf Kontoebene für Entwicklungs-, Test- und Produktionsworkloads aus.
- Streamen Sie ML-Workload-Protokolle in ein Protokollarchivkonto und filtern Sie anschließend die Protokollanalyse in einem Observability-Konto und wenden Sie sie an.
- Führen Sie ein zentrales Governance-Konto für die Bereitstellung, Steuerung und Prüfung des Datenzugriffs ein.
- Integrieren Sie Sicherheits- und Governance-Dienste (SGS) mit entsprechenden präventiven und detektiven Schutzmaßnahmen in jedes Konto, um Sicherheit und Compliance gemäß Ihren Unternehmens- und Workload-Anforderungen zu gewährleisten.

Zentralisierte Modellkontenstruktur

In diesem Modell ist das ML-Plattformteam verantwortlich für die Bereitstellung von:

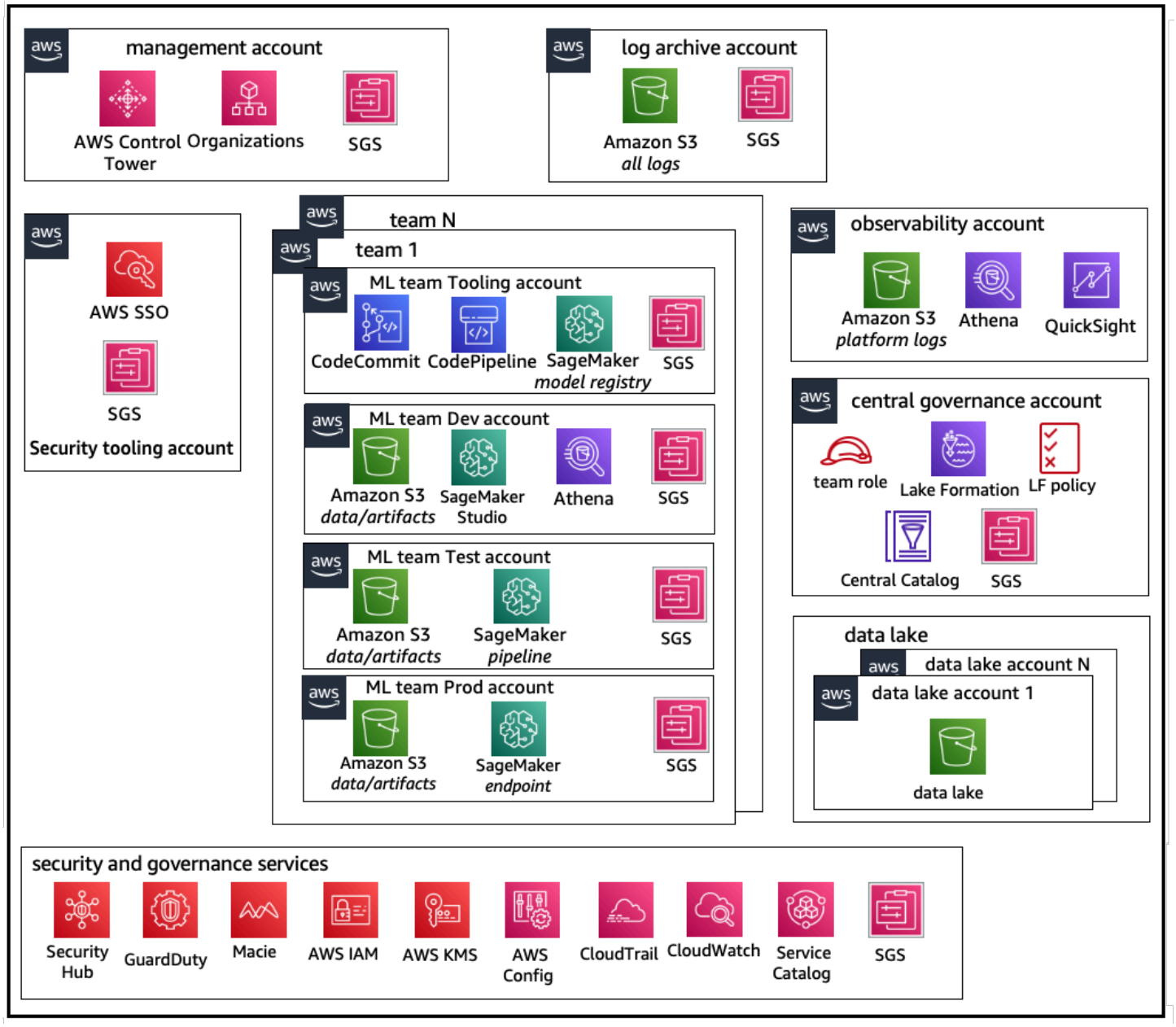
- Ein Shared-Services-Tooling-Konto, das die Anforderungen von Machine Learning Operations (MLOps) in allen Data-Science-Teams erfüllt.
- Konten für die Entwicklung, den Test und die Produktion von ML-Workloads, die von allen Data-Science-Teams gemeinsam genutzt werden.
- Governance-Richtlinien, um sicherzustellen, dass die Workloads jedes Data-Science-Teams isoliert ausgeführt werden.
- Allgemeine bewährte Verfahren.



Kontostruktur eines zentralisierten Betriebsmodells

Dezentrale Modellkontenstruktur

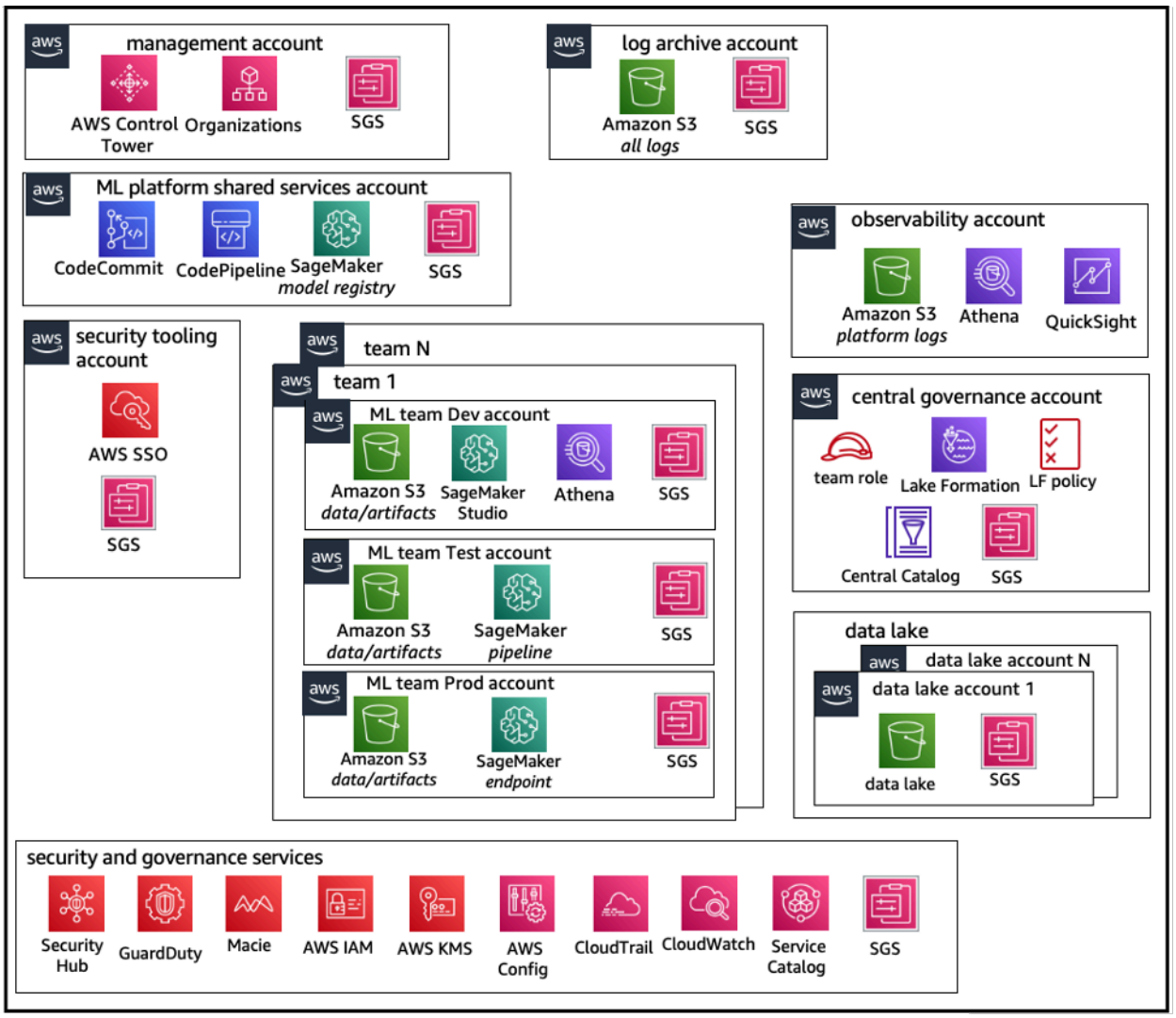
In diesem Modell arbeitet jedes ML-Team unabhängig bei der Bereitstellung, Verwaltung und Verwaltung von ML-Konten und -Ressourcen. Wir empfehlen ML-Teams jedoch, einen zentralisierten Ansatz für Beobachtbarkeit und Datenverwaltung zu verwenden, um die Datenverwaltung und das Auditmanagement zu vereinfachen.



Kontostruktur eines dezentralen Betriebsmodells

Kontostruktur nach föderiertem Modell

Dieses Modell ähnelt dem zentralisierten Modell. Der Hauptunterschied besteht jedoch darin, dass jeder science/ML team gets their own set of development/test/production Daten-Workload eine robuste physische Isolierung seiner ML-Ressourcen ermöglicht und es jedem Team ermöglicht, unabhängig zu skalieren, ohne andere Teams zu beeinträchtigen.



Kontostruktur eines föderierten Betriebsmodells

Mehrmandantenfähigkeit der ML-Plattform

Multitenancy ist eine Softwarearchitektur, bei der eine einzelne Softwareinstanz mehrere unterschiedliche Benutzergruppen bedienen kann. Ein Mandant ist eine Gruppe von Benutzern, die gemeinsamen Zugriff mit bestimmten Rechten auf die Softwareinstanz haben. Wenn Sie beispielsweise mehrere ML-Produkte entwickeln, kann jedes Produktteam mit ähnlichen Zugriffsanforderungen als Mandant oder Team betrachtet werden.

Es ist zwar möglich, mehrere Teams innerhalb einer SageMaker AI Studio-Instanz (z. B. SageMaker AI Domain) zu implementieren, aber wägen Sie diese Vorteile gegen Kompromisse wie Explosionsradius, Kostenzuweisung und Konto-Level-Beschränkungen ab, wenn Sie mehrere Teams

in einer einzigen SageMaker AI Studio-Domain zusammenführen. In den folgenden Abschnitten erfahren Sie mehr über diese Kompromisse und Best Practices.

Wenn Sie eine absolute Ressourcenisolierung benötigen, sollten Sie die Implementierung von SageMaker AI Studio-Domänen für jeden Mandanten in einem anderen Konto in Betracht ziehen. Abhängig von Ihren Isolationsanforderungen können Sie mehrere Geschäftsbereiche (LOBs) als mehrere Domänen innerhalb eines einzigen Kontos und einer Region implementieren. Nutzen Sie gemeinsam genutzte Bereiche für die Zusammenarbeit zwischen Mitgliedern desselben Teams/ LOB nahezu in Echtzeit. Bei mehreren Domänen verwenden Sie weiterhin Richtlinien und Berechtigungen für Identity Access Management (IAM), um die Isolierung von Ressourcen sicherzustellen.

SageMaker KI-Ressourcen, die aus einer Domain erstellt wurden, werden automatisch mit dem Amazon-Ressourcennamen der Domain (ARN) und dem Benutzerprofil oder Bereich versehen, ARN um Ressourcen einfach zu isolieren. Beispielrichtlinien finden Sie in der Dokumentation zur Isolierung von Domänenressourcen. Dort finden Sie die ausführliche Referenz, wann eine Strategie mit mehreren Konten oder mehreren Domänen verwendet werden sollte, sowie die Funktionsvergleiche in der Dokumentation. Außerdem können Sie sich Beispielskripts ansehen, um Tags für bestehende Domänen im Repository aufzufüllen. GitHub

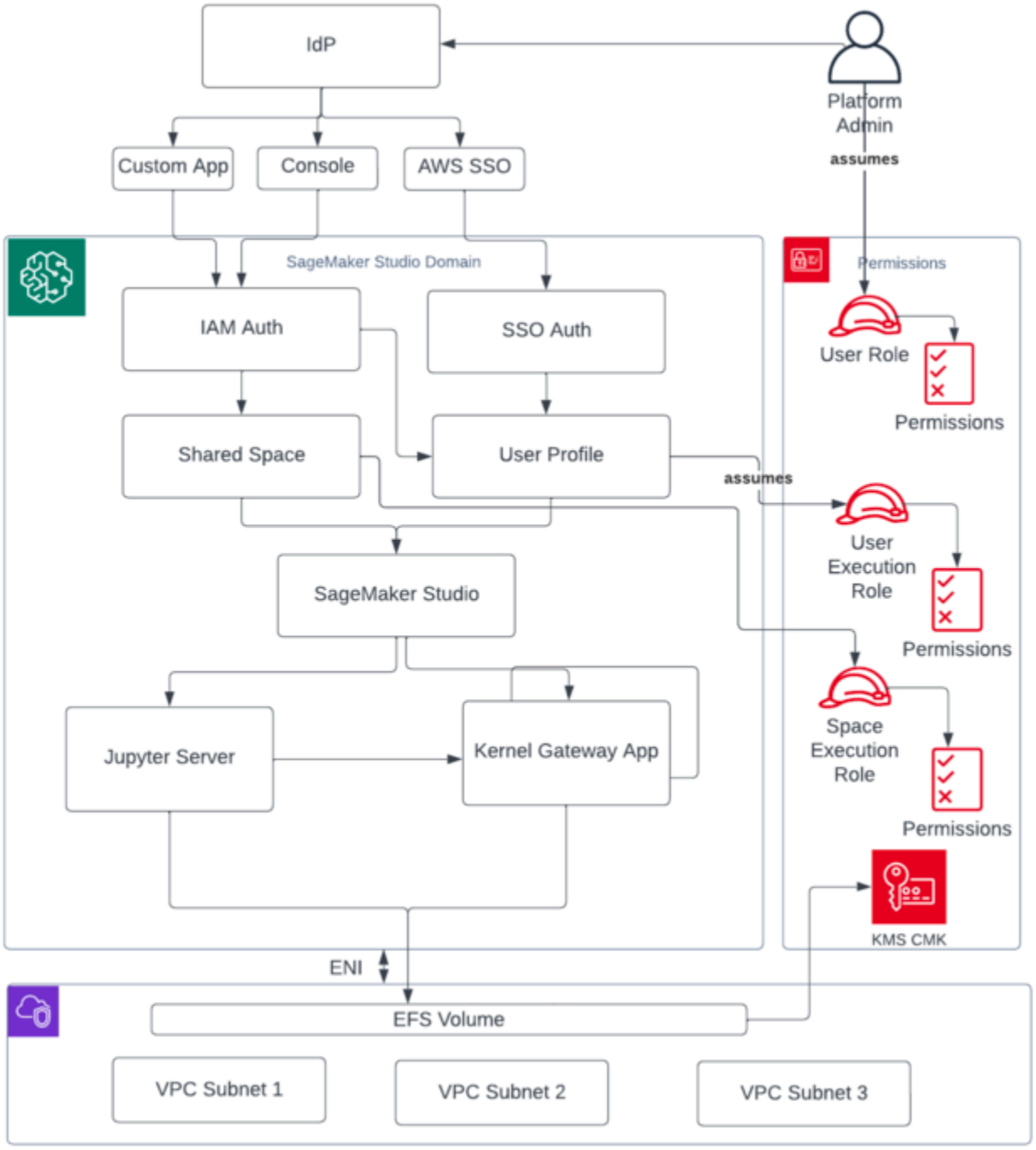
Schließlich können Sie mit Hilfe von AI Studio-Ressourcen eine Self-Service-Bereitstellung von SageMaker AI Studio-Ressourcen für mehrere Konten implementieren. AWS Service Catalog Weitere Informationen finden Sie unter AWS Service Catalog Produkte in mehreren AWS-Konten und AWS-Regionen verwalten.

Domänenverwaltung

Eine Amazon SageMaker AI-Domain besteht aus:

- Ein zugeordnetes Amazon Elastic File System (AmazonEFS) -Volume
- Eine Liste autorisierter Benutzer
- Eine Vielzahl von Sicherheits-, Anwendungs-, Richtlinien- und Amazon Virtual Private Cloud (AmazonVPC) -Konfigurationen

Das folgende Diagramm bietet einen allgemeinen Überblick über die verschiedenen Komponenten, die eine SageMaker AIStudio Domain ausmachen:

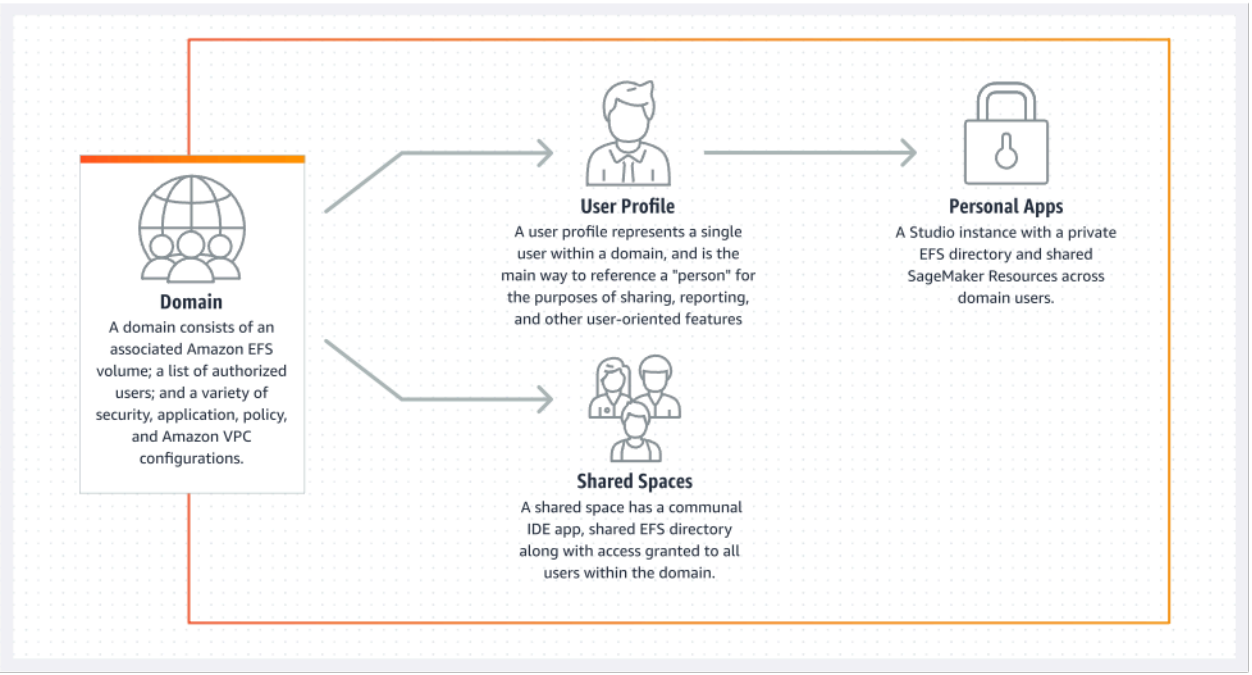


Überblick über verschiedene Komponenten, die eine SageMaker AI Studio-Domäne bilden

Mehrere Domains und gemeinsam genutzte Bereiche

Amazon SageMaker AI unterstützt jetzt die Erstellung mehrerer SageMaker KI-Domains in einer einzigen AWS-Region für jedes Konto. Jede Domain kann ihre eigenen Domain-Einstellungen wie den Authentifizierungsmodus und Netzwerkeinstellungen wie VPC Subnetze haben. Ein Benutzerprofil kann nicht domänenübergreifend gemeinsam genutzt werden. Wenn ein menschlicher Benutzer Teil mehrerer Teams ist, die durch Domänen getrennt sind, erstellen Sie in jeder Domäne ein Benutzerprofil für den Benutzer. Weitere Informationen zum Hinterfüllen von Tags für bestehende Domänen finden Sie in der Übersicht über mehrere Domänen.

Jede im IAM Authentifizierungsmodus eingerichtete Domain kann gemeinsam genutzten Speicherplatz für die Zusammenarbeit zwischen Benutzern nahezu in Echtzeit nutzen. Mit einem gemeinsamen Bereich erhalten Benutzer Zugriff auf ein gemeinsames EFS Amazon-Verzeichnis und eine gemeinsam genutzte JupyterServerApp für die Benutzeroberfläche und können sie nahezu in Echtzeit gemeinsam bearbeiten. Die automatische Kennzeichnung von Ressourcen, die in gemeinsam genutzten Bereichen erstellt wurden, ermöglicht es den Administratoren, die Kosten auf Projektebene zu verfolgen. Die gemeinsam genutzte JupyterServer Benutzeroberfläche filtert auch Ressourcen wie Experimente und Modellregistrierungseinträge, sodass nur Elemente angezeigt werden, die für das gemeinsame ML-Projekt relevant sind. Das folgende Diagramm bietet einen Überblick über private Apps und gemeinsam genutzte Bereiche innerhalb der einzelnen Domänen.



Überblick über private Apps und gemeinsam genutzte Bereiche innerhalb einer einzigen Domain

Richten Sie gemeinsame Bereiche in Ihrer Domain ein

Gemeinsam genutzte Bereiche werden in der Regel für ein bestimmtes ML-Unterfangen oder -Projekt erstellt, bei dem Mitglieder einer einzelnen Domain nahezu in Echtzeit Zugriff auf denselben zugrunde liegenden Dateispeicher und IDE benötigen. Der Benutzer kann nahezu in Echtzeit auf seine Notizbücher zugreifen, sie lesen, bearbeiten und teilen, was ihm den schnellsten Weg bietet, mit seinen Kollegen zu iterieren.

Um einen gemeinsam genutzten Bereich zu erstellen, müssen Sie zunächst eine standardmäßige Ausführungsrolle für den Bereich festlegen, die die Berechtigungen aller Benutzer bestimmt, die den Bereich nutzen. Zum Zeitpunkt der Erstellung dieses Artikels haben alle Benutzer innerhalb einer Domäne Zugriff auf alle gemeinsam genutzten Bereiche in ihrer Domäne. Die aktuelle Dokumentation zum Hinzufügen von Shared Spaces zu einer bestehenden Domain finden Sie unter Shared Space erstellen.

Richten Sie Ihre Domain für den IAM Verbund ein

Bevor Sie den AWS Identity and Access Management (IAM) -Verbund für Ihre SageMaker AI Studio-Domain einrichten, müssen Sie eine IAM Verbundbenutzerrolle (z. B. einen Plattformadministrator) in Ihrem IdP einrichten, wie im Abschnitt Identitätsmanagement beschrieben.

Detaillierte Anweisungen zur Einrichtung von SageMaker AI Studio mit dieser IAM Option finden Sie unter Onboard to Amazon SageMaker Domain Using IAM Identity Center.

Richten Sie Ihre Domain für den Single Sign-On (SSO) -Verbund ein

Um den Single Sign-On (SSO) -Verbund zu verwenden, müssen Sie AWS IAM Identity Center ihn in Ihrem AWS OrganizationsVerwaltungskonto in derselben Region aktivieren, in der Sie SageMaker AI Studio ausführen müssen. Die Schritte zur Einrichtung der Domäne ähneln den Schritten für den IAM Verbund, mit der Ausnahme, dass Sie im Abschnitt Authentifizierung die Option AWS IAM Identity Center(iDC) auswählen.

Eine ausführliche Anleitung finden Sie unter Onboarding to Amazon SageMaker Domain Using IAM Identity Center.

SageMaker AI Studio-Benutzerprofil

Ein Benutzerprofil stellt einen einzelnen Benutzer innerhalb einer Domain dar und ist die wichtigste Methode, um auf eine „Person“ Bezug zu nehmen, um sie zu teilen, Berichte zu erstellen und andere benutzerorientierte Funktionen zu nutzen. Diese Entität wird erstellt, wenn ein Benutzer toSageMaker AI Studio einloggt. Wenn ein Administrator eine Person per E-Mail einlädt oder sie aus IdC importiert, wird automatisch ein Benutzerprofil erstellt. Ein Benutzerprofil ist der primäre Inhaber der Einstellungen für einen einzelnen Benutzer und enthält einen Verweis auf das private Amazon Elastic File System (AmazonEFS) -Home-Verzeichnis des Benutzers. Wir empfehlen, für jeden physischen Benutzer der SageMaker AI Studio-Anwendung ein Benutzerprofil zu erstellen. Jeder Benutzer hat sein eigenes Verzeichnis bei AmazonEFS, und Benutzerprofile können nicht domänenübergreifend in demselben Konto gemeinsam genutzt werden.

Jedes Benutzerprofil, das die SageMaker AI Studio-Domain teilt, erhält dedizierte Rechenressource (n) (z. B. SageMaker AI Amazon Elastic Compute Cloud (AmazonEC2) -Instanz (en)) zum Ausführen von Notebooks. Die Compute-Instances, die Benutzer eins zugewiesen sind, sind vollständig von denen isoliert, die Benutzer zwei zugewiesen sind. In ähnlicher Weise sind die Rechenressourcen, die Benutzern in einem AWS Konto zugewiesen sind, vollständig von denen getrennt, die Benutzern in einem anderen Konto zugewiesen sind. Jeder Benutzer kann bis zu vier Anwendungen (Apps) in isolierten Docker-Containern oder Images auf demselben Instanztyp ausführen.

Jupyter Server-App

Wenn Sie ein Amazon SageMaker AI Studio-Notizbuch für einen Benutzer starten, indem Sie auf das vorsignierte Notizbuch zugreifen URL oder sich mit AWS IAM iDC anmelden, wird die Jupyter Server-App in der vom AI-Service verwalteten Instance gestartet. SageMaker VPC Jeder Benutzer erhält seine eigene dedizierte Jupyter Server-App in einer privaten App. Standardmäßig wird die Jupyter Server-App für SageMaker AI Studio-Notebooks auf einer dedizierten *ml.t3.medium* Instanz ausgeführt (die als Systeminstanztyp reserviert ist). Die Rechenleistung für diese Instanz wird dem Kunden nicht in Rechnung gestellt.

Die Jupyter Kernel Gateway-App

Die Kernel Gateway-App kann über die API oder die SageMaker AI Studio-Oberfläche erstellt werden und läuft auf dem ausgewählten Instanztyp. Diese App kann mit einem der integrierten SageMaker AI Studio-Images ausgeführt werden, die mit gängigen Datenwissenschaft- und Deep-Learning-Paketen wie TensorFlowApache MXNet und PyTorchvorkonfiguriert sind.

Benutzer können mehrere Jupyter-Notebook-Kernel, Terminalsitzungen und interaktive Konsolen im selben Studio starten und ausführen. 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

Um zusätzliche Apps zu erstellen, müssen Sie einen anderen Instanztyp verwenden. In einem Benutzerprofil kann nur eine Instanz eines beliebigen Instanztyps ausgeführt werden. Beispielsweise kann ein Benutzer sowohl ein einfaches Notebook mit dem integrierten Data-Science-Image von SageMaker AI Studio als auch ein anderes Notebook mit dem integrierten TensorFlow Image auf derselben Instanz ausführen. Benutzern wird die Zeit in Rechnung gestellt, in der die Instanz ausgeführt wird. Um Kosten zu vermeiden, wenn der Benutzer SageMaker AI Studio nicht aktiv ausführt, muss der Benutzer die Instanz herunterfahren. Weitere Informationen finden Sie unter Studio-Apps herunterfahren und aktualisieren.

Jedes Mal, wenn Sie eine Kernel Gateway-App über die SageMaker AI Studio-Oberfläche herunterfahren und erneut öffnen, wird diese App auf einer neuen Instanz gestartet. Das bedeutet, dass die Installation des Pakets nicht durch Neustarts derselben App beibehalten wird. Ebenso gehen die installierten Pakete und Sitzungsvariablen verloren, wenn ein Benutzer den Instanztyp auf einem Notebook ändert. Sie können jedoch Funktionen wie Bring Your Own Image und Lifecycle-Skripte verwenden, um die eigenen Pakete des Benutzers in SageMaker AI Studio zu übertragen und sie über Instanzwechsel und den Start neuer Instanzen beizubehalten.

Amazon Elastic File System-Volume

Wenn eine Domain erstellt wird, wird ein einzelnes Amazon Elastic File System (AmazonEFS) - Volume erstellt, das von allen Benutzern innerhalb der Domain verwendet werden kann. Jedes Benutzerprofil erhält ein privates Home-Verzeichnis innerhalb des EFS Amazon-Volumes, in dem die Notizbücher, GitHub Repositorys und Datendateien des Benutzers gespeichert werden. Jeder Bereich innerhalb einer Domain erhält ein privates Verzeichnis innerhalb des EFS Amazon-Volumes, auf das mehrere Benutzerprofile zugreifen können. Der Zugriff auf die Ordner ist durch Dateisystemberechtigungen nach Benutzern getrennt. SageMaker AI Studio erstellt für jedes Benutzerprofil oder jeden Bereich eine globale eindeutige Benutzer-ID und wendet diese als tragbare Betriebssystemschnittstelle (POSIX) für den Zugriff auf die Daten user/group ID for the user's home directory on EFS, which prevents other users/spaces an.

Sicherung und Wiederherstellung

Ein vorhandenes EFS Volume kann nicht an eine neue SageMaker AI-Domain angehängt werden. Stellen Sie in einer Produktionsumgebung sicher, dass das EFS Amazon-Volume gesichert ist (auf einem anderen EFS Volume oder auf Amazon Simple Storage Service (Amazon S3)). Wenn ein EFS Volume versehentlich gelöscht wird, muss der Administrator die SageMaker AI Studio-Domain entfernen und neu erstellen. Der Prozess läuft folgendermaßen ab:

Erstellen Sie über die `DescribeSpace` API Aufrufe,, und eine Sicherungskopie der Liste der EFS Benutzerprofile `ListUserProfiles` `DescribeUserProfileList Spaces`, Bereiche und des zugehörigen Benutzers IDs (UIDs).

1. Erstellen Sie eine neue SageMaker AI Studio-Domäne.
2. Erstellen Sie die Benutzerprofile und Bereiche.
3. Kopieren Sie für jedes Benutzerprofil die Dateien aus dem Backup auf EFS /Amazon S3.
4. Löschen Sie optional alle Apps und Benutzerprofile in der alten SageMaker AI Studio-Domain.

Detaillierte Anweisungen finden Sie im Anhang, Abschnitt SageMaker AI Studio-Domain-Backup und -Wiederherstellung.

> **Note**
>
> Dies kann auch dadurch erreicht werden`LifecycleConfigurations`, dass jedes Mal, wenn ein Benutzer seine App startet, Daten auf und von S3 gesichert werden.

EBSAmazon-Volumen

Jeder SageMaker AI Studio Notebook-Instance ist außerdem ein Amazon Elastic Block Store (AmazonEBS) -Speichervolume zugeordnet. Es wird als Root-Volume des Containers oder Images verwendet, das auf der Instance ausgeführt wird. Während der EFS Amazon-Speicher persistent ist, ist das an den Container angehängte EBS Amazon-Volume temporär. Die lokal auf Amazon EBS Volume gespeicherten Daten werden nicht dauerhaft gespeichert, wenn der Kunde die App löscht.

Sicherung des Zugriffs auf die vorab signierte Datei URL

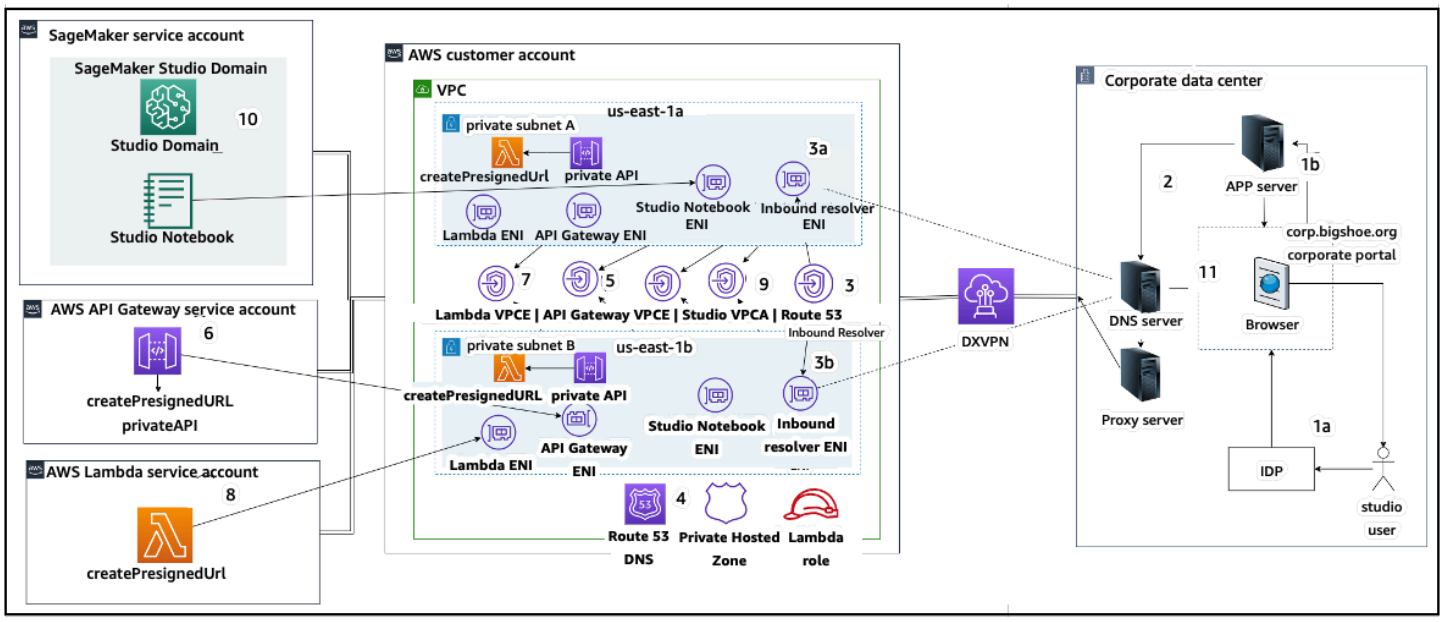
Wenn ein SageMaker AI Studio-Benutzer den Notizbuch-Link öffnet, validiert SageMaker AI Studio die IAM Richtlinie des Verbundbenutzers zur Autorisierung des Zugriffs und generiert die vorsignierte Version für den Benutzer und löst sie auf. URL Da die SageMaker AI-Konsole auf einer Internetdomäne läuft, URL ist diese generierte, vorsignierte Datei in der Browsersitzung sichtbar. Dies stellt einen unerwünschten Bedrohungsvektor für Datendiebstahl und den Zugriff auf Kundendaten dar, wenn keine angemessenen Zugriffskontrollen durchgesetzt werden.

Studio unterstützt einige Methoden zur Durchsetzung von Zugriffskontrollen gegen URL vorsignierten Datendiebstahl:

- IP-Validierung des Clients anhand der Richtlinienbedingung IAM `aws:sourceIp`
- VPCClient-Validierung anhand der IAM Bedingung `aws:sourceVpc`
- Validierung des VPC Client-Endpunkts mithilfe der IAM Richtlinienbedingung `aws:sourceVpce`

Wenn Sie von der AI-Konsole aus auf SageMaker SageMaker AI Studio-Notebooks zugreifen, besteht die einzige verfügbare Option darin, die Client-IP-Validierung mit der IAM Richtlinienbedingung zu verwenden`aws:sourceIp`. Sie können jedoch Produkte zum Routing von Browser-Traffic wie Zscaler verwenden, um sicherzustellen, dass der Internetzugang Ihrer Belegschaft skalierbar und gesetzeskonform ist. Diese Traffic-Routing-Produkte generieren ihre eigene Quell-IP, deren IP-Bereich nicht vom Unternehmenskunden kontrolliert wird. Dies macht es diesen Unternehmenskunden unmöglich, die `aws:sourceIp` Bedingung zu nutzen.

Um die VPC Client-Endpunktvalidierung anhand der IAM Richtlinienbedingung zu verwenden`aws:sourceVpce`, URL muss die Erstellung einer vorab signierten Anforderung von demselben Kunden stammen, auf VPC dem SageMaker AI Studio bereitgestellt wird, und die Lösung der vorab signierten URL Anforderungen muss über einen SageMaker AI VPC Studio-Endpunkt auf dem Kunden erfolgen. VPC Diese Auflösung der vorsignierten Daten URL während des Zugriffs für Benutzer des Unternehmensnetzwerks kann mithilfe von DNS Weiterleitungsregeln (sowohl in Zscaler als auch in CorporateDNS) und dann mithilfe eines Amazon Route 53-Inbound-Resolvers zum VPC Kundenendpunkt erfolgen, wie in der folgenden Architektur dargestellt:



Zugriff auf Studio URL mit vorsigniertem VPC Endpunkt über das Unternehmensnetzwerk

step-by-stepAnleitungen zur Einrichtung der vorherigen Architektur finden Sie unter Secure Amazon SageMaker AI Studio Presigned URLs Part 1: Fundamentale Infrastruktur.

SageMaker Kontingente und Limits für KI-Domains

- SageMaker Der AI SSO Studio-Domänenverbund wird nur in der Region unterstützt, und zwar für alle Mitgliedskonten der AWS Organisation, in der AWS Identity Center bereitgestellt wird.
- Gemeinsam genutzte Bereiche werden derzeit nicht für Domains unterstützt, die mit AWS Identity Center eingerichtet wurden.
- VPCund die Subnetzkonfiguration kann nach der Erstellung der Domain nicht geändert werden. Sie können jedoch eine neue Domäne mit einer anderen VPC Subnetzkonfiguration erstellen.
- Der Domänenzugriff kann nach dem Erstellen der Domäne nicht zwischen den SSO Modi IAM und geändert werden. Sie können eine neue Domain mit einem anderen Authentifizierungsmodus erstellen.
- Es gibt ein Limit von vier Kernel-Gateway-Apps pro Instance-Typ, die für jeden Benutzer gestartet werden.
- Jeder Benutzer kann nur eine Instanz jedes Instanztyps starten.
- Es gibt Beschränkungen für den Ressourcenverbrauch innerhalb einer Domain, z. B. die Anzahl der nach Instance-Typen gestarteten Instanzen und die Anzahl der Benutzerprofile, die erstellt

werden können. Eine vollständige Liste der Servicebeschränkungen finden Sie auf der Seite mit den Servicekontingenten.

- Kunden können eine Support-Anfrage mit geschäftlicher Begründung einreichen, um die standardmäßigen Ressourcenlimits, wie z. B. die Anzahl der Domänen oder Benutzerprofile, zu erhöhen, für die Einschränkungen auf Kontoebene gelten.
- Das feste Limit für die Anzahl gleichzeitiger Apps pro Konto liegt bei 2.500 Apps. Die Beschränkungen für Domänen und Benutzerprofile hängen von diesem festen Limit ab. Ein Konto kann beispielsweise eine einzelne Domäne mit 1.000 Benutzerprofilen oder 20 Domänen mit jeweils 50 Benutzerprofilen haben.

Identitätsverwaltung

In diesem Abschnitt wird erläutert, wie sich Workforce-Benutzer in einem Unternehmensverzeichnis zu SageMaker AI Studio zusammenschließen AWS-Konten und darauf zugreifen. Zunächst beschreiben wir kurz, wie Benutzer, Gruppen und Rollen zugeordnet werden und wie der Benutzerverbund funktioniert.

Benutzer, Gruppen und Rollen

AWS In werden Ressourcenberechtigungen mithilfe von Benutzern, Gruppen und Rollen verwaltet. Kunden können ihre Benutzer und Gruppen entweder über IAM oder in einem Unternehmensverzeichnis wie Active Directory (AD) verwalten, das über einen externen IdP wie Okta aktiviert wird, sodass sie die Benutzer für verschiedene Anwendungen authentifizieren können, die in der Cloud und vor Ort ausgeführt werden.

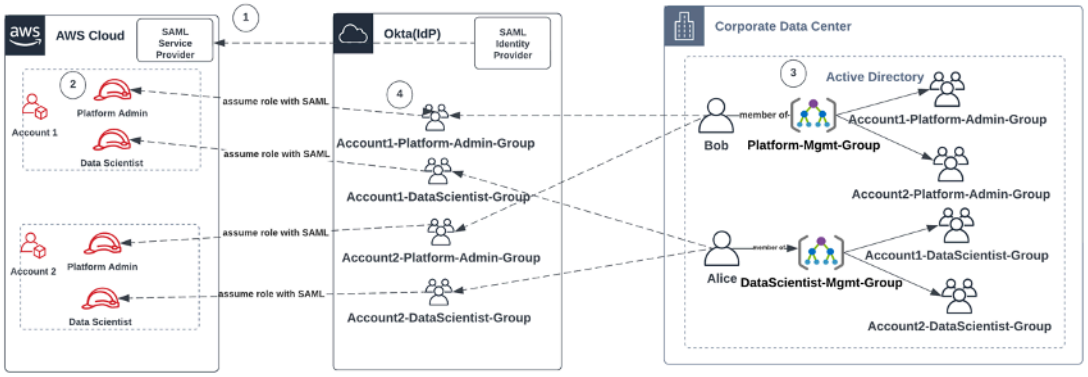
Wie im Abschnitt AWS Security Pillar Identity Management beschrieben, ist es eine bewährte Methode, Ihre Benutzeridentitäten in einem zentralen IdP zu verwalten, da dies die einfache Integration in Ihre Backend-HR-Prozesse erleichtert und Ihnen hilft, den Zugriff auf Ihre Belegschaftsbenutzer zu verwalten.

IdPs wie Okta ermöglicht es Endbenutzern, sich bei einer oder mehreren zu authentifizieren AWS-Konten und mithilfe SSO der Security Assertation Markup Language () Zugriff auf bestimmte Rollen zu erhalten. SAML IdP-Administratoren haben die Möglichkeit, Rollen aus dem AWS-Konten IdP herunterzuladen und diese Benutzern zuzuweisen. Bei der Anmeldung wird Endbenutzern ein AWS Bildschirm mit einer Liste der Rollen angezeigt, die ihnen in einer oder mehreren AWS Rollen zugewiesen wurden. AWS AWS-Konten Sie können die Rolle auswählen, die sie bei der Anmeldung übernehmen möchten, wodurch ihre Berechtigungen für die Dauer der authentifizierten Sitzung definiert werden.

Für jede Kombination aus Konto und Rolle, auf die Sie Zugriff gewähren möchten, muss in IdP eine Gruppe vorhanden sein. Sie können sich diese Gruppen als AWS rollenspezifische Gruppen vorstellen. Jedem Benutzer, der Mitglied dieser rollenspezifischen Gruppen ist, wird eine einzige Berechtigung gewährt: Zugriff auf eine bestimmte Rolle in einer bestimmten. AWS-Konto Dieser Prozess mit einer einzigen Berechtigung lässt sich jedoch nicht auf die Verwaltung des Benutzerzugriffs skalieren, indem jeder Benutzer bestimmten AWS Rollengruppen zugewiesen wird. Um die Verwaltung zu vereinfachen, empfehlen wir Ihnen außerdem, eine Reihe von Gruppen für

alle unterschiedlichen Benutzergruppen in Ihrer Organisation zu erstellen, für die unterschiedliche Berechtigungssätze erforderlich sind. AWS

Um die zentrale IdP-Einrichtung zu veranschaulichen, stellen Sie sich ein Unternehmen mit AD-Setup vor, in dem Benutzer und Gruppen mit dem IdP-Verzeichnis synchronisiert werden. AWS In sind diese AD-Gruppen Rollen zugeordnet. IAM Die wichtigsten Schritte des Workflows sind wie folgt:



Workflow für das Onboarding von AD-Benutzern, AD-Gruppen und -Rollen IAM

1. In AWS, Richten Sie die SAML Integration für jeden von Ihnen AWS-Konten mit Ihrem IdP ein.
2. AWS Richten Sie in jedem von ihnen Rollen ein AWS-Konto und synchronisieren Sie sie mit dem IdP.
3. Im AD-System des Unternehmens:
   a. Erstellen Sie eine AD-Gruppe für jede Kontorolle und synchronisieren Sie sie mit dem IdP `Account1-Platform-Admin-Group` (z. B. AWS Rollengruppe).
   b. Erstellen Sie eine Verwaltungsgruppe auf jeder Persona-Ebene (z. B.`Platform-Mgmt-Group`) und weisen Sie AWS Rollengruppen als Mitglieder zu.
   c. Weisen Sie dieser Verwaltungsgruppe Benutzer zu, um Zugriff auf AWS-Konto Rollen zu gewähren.
4. Ordnen Sie in IdP AWS Rollengruppen (z. B.`Account1-Platform-Admin-Group`) AWS-Konto Rollen zu (z. B. Platform Admin in Account1).
5. Wenn sich Data Scientist Alice bei Idp anmeldet, wird ihr eine AWS Federation App-Benutzeroberfläche mit zwei Optionen zur Auswahl angezeigt: „Account 1 Data Scientist“ und „Account 2 Data Scientist“.
6. Alice wählt die Option „Account 1 Data Scientist“ und sie werden mit ihrer autorisierten Anwendung in Konto 1 (AI Console) verbunden. AWS SageMaker

Eine ausführliche Anleitung zur Einrichtung eines SAML Kontoverbunds finden Sie in Oktas How to Configure SAML 2.0 for Account Federation. AWS

Benutzerverbund

Die Authentifizierung für SageMaker AI Studio kann entweder mit IAM oder IAM iDC erfolgen. Wenn die Benutzer verwaltet werdenIAM, können sie den IAM Modus wählen. Wenn das Unternehmen einen externen IdP verwendet, kann es sich entweder über IAM oder IAM über IdC zusammenschließen. Beachten Sie, dass der Authentifizierungsmodus für eine bestehende SageMaker AI Studio-Domäne nicht aktualisiert werden kann. Daher ist es wichtig, die Entscheidung zu treffen, bevor Sie eine SageMaker AI Studio-Produktionsdomäne erstellen.

Wenn SageMaker AI Studio im IAM Modus eingerichtet ist, greifen SageMaker AI Studio-Benutzer über eine vorsignierte App auf die App zuURL, die einen Benutzer automatisch bei der SageMaker AI Studio-App anmeldet, wenn sie über einen Browser darauf zugreifen.

IAM-Benutzer

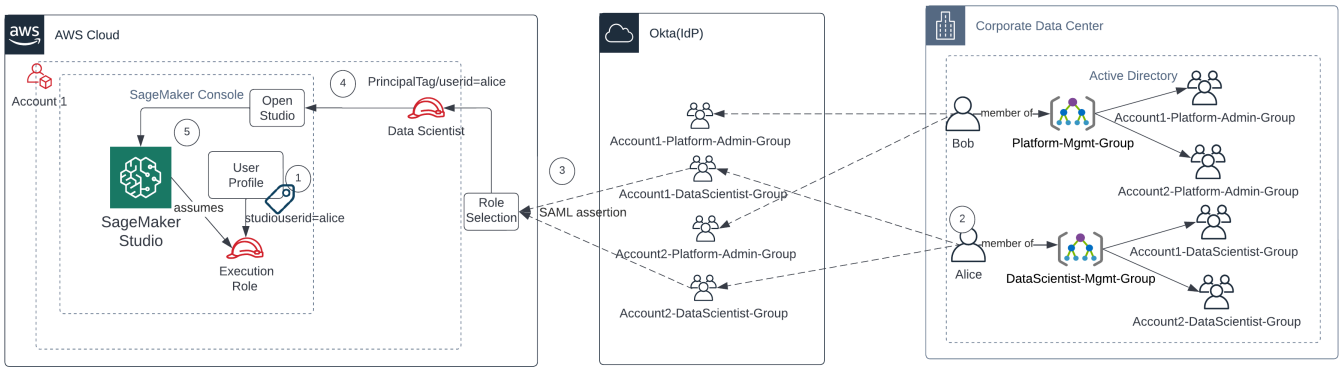
Für IAM Benutzer erstellt der Administrator SageMaker AI Studio-Benutzerprofile für jeden Benutzer und ordnet das Benutzerprofil einer IAM Rolle zu, die die erforderlichen Aktionen ermöglicht, die der Benutzer in Studio ausführen muss. Um zu verhindern, dass ein AWS Benutzer nur auf sein SageMaker AI Studio-Benutzerprofil zugreift, sollte der Administrator das SageMaker AI Studio-Benutzerprofil taggen und dem Benutzer eine IAM Richtlinie hinzufügen, die ihm nur Zugriff gewährt, wenn der Tagwert mit dem AWS Benutzernamen identisch ist. Die Grundsatzerklärung sieht wie folgt aus:

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

```
            }
        }
    ]
}
```

AWS IAModer Kontoverbund

Die AWS-Konto Verbundmethode ermöglicht es Kunden, sich von ihrem SAML IdP wie Okta aus mit der SageMaker KI-Konsole zu verbinden. Um zu verhindern, dass Benutzer nur auf ihr Benutzerprofil zugreifen, sollte der Administrator das SageMaker AI Studio-Benutzerprofil taggen, `PrincipalTags` den IdP hinzufügen und sie als transitive Tags festlegen. Das folgende Diagramm zeigt, wie der Verbundbenutzer (Data Scientist Alice) autorisiert ist, auf sein eigenes SageMaker AI Studio-Benutzerprofil zuzugreifen.



Zugriff auf SageMaker AI Studio im IAM Verbundmodus

1. Das Alice SageMaker AI Studio-Benutzerprofil ist mit seiner Benutzer-ID gekennzeichnet und der Ausführungsrolle zugeordnet.
2. Alice authentifiziert sich bei IdP (Okta).
3. IdP authentifiziert Alice und veröffentlicht eine SAML Assertion mit den beiden Rollen (Data Scientist für Konten 1 und 2), bei denen Alice Mitglied ist. Alice wählt die Rolle Data Scientist für Konto 1 aus.
4. Alice ist bei Account 1 SageMaker AI Console angemeldet und hat die Rolle des Datenwissenschaftlers übernommen. Alice öffnet ihre Studio-App-Instanz aus der Liste der Studio-App-Instanzen.
5. Das Alice-Prinzipal-Tag in der angenommenen Rollensitzung wird anhand des Benutzerprofil-Tags der ausgewählten SageMaker AI Studio-App-Instanz validiert. Wenn das Profil-Tag gültig ist, wird die SageMaker AI Studio-App-Instanz gestartet, wobei die Ausführungsrolle übernommen wird.

Wenn Sie die Erstellung von Rollen und Richtlinien für die SageMaker KI-Ausführung im Rahmen des Benutzer-Onboardings automatisieren möchten, können Sie dies wie folgt erreichen:

1. Richten Sie eine AD-Gruppe ein, z. B. `SageMaker AI-Account1-Group` auf Konto- und Studio-Domänenebene.
2. Fügen Sie SageMaker AI-Account1-Group zur Gruppenmitgliedschaft des Benutzers hinzu, wenn Sie einen Benutzer in AI Studio einbinden müssen. SageMaker

Richten Sie einen Automatisierungsprozess ein, der das `SageMaker AI-Account1-Group` Mitgliedsereignis überwacht und anhand AWS APIs dessen die Rolle, die Richtlinien, die Tags und das SageMaker AI Studio-Benutzerprofil auf der Grundlage der AD-Gruppenmitgliedschaften erstellt werden. Ordnen Sie die Rolle dem Benutzerprofil zu. Eine Beispielrichtlinie finden Sie unterVerhindern Sie, dass SageMaker AI Studio-Benutzer auf andere Benutzerprofile zugreifen.

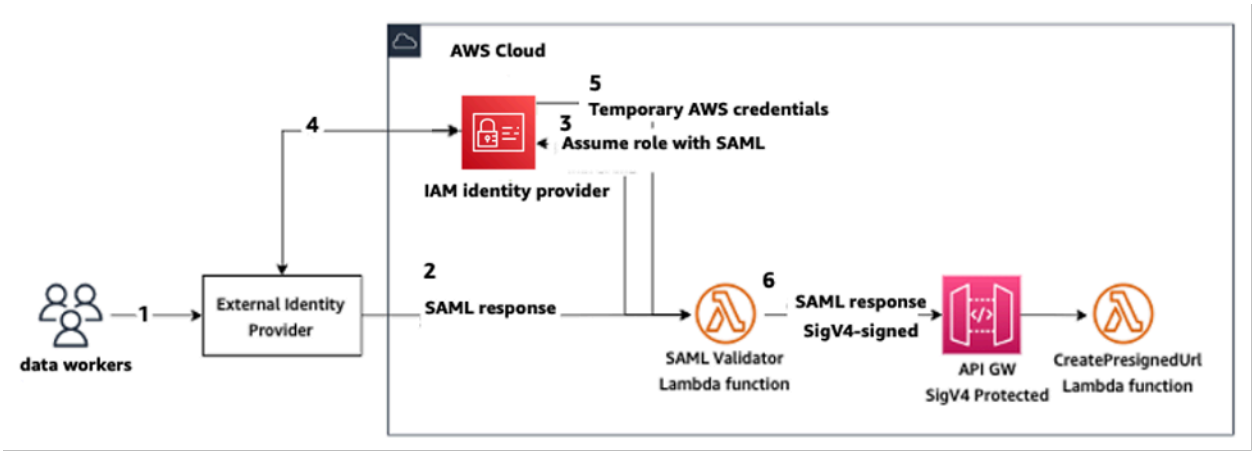
SAMLAuthentifizierung mit AWS Lambda

Im IAM Modus können Benutzer auch mithilfe von SAML Assertionen in SageMaker AI Studio authentifiziert werden. In dieser Architektur verfügt der Kunde über einen vorhandenen IdP, über den er eine SAML Anwendung erstellen kann, mit der die Benutzer auf Studio zugreifen können (anstelle der AWS Identity Federation-Anwendung). Der IdP des Kunden wird hinzugefügtIAM. Eine AWS Lambda Funktion hilft bei der Validierung der SAML Assertion mithilfe von IAM und STS und ruft dann direkt ein API Gateway oder eine Lambda-Funktion auf, um die vorsignierte Domain zu erstellen. URL

Der Vorteil dieser Lösung besteht darin, dass die Lambda-Funktion die Logik für den Zugriff auf SageMaker AI Studio anpassen kann. Beispielsweise:

- Erstellen Sie automatisch ein Benutzerprofil, falls noch keines vorhanden ist.
- Hängen Sie Rollen oder Richtliniendokumente an die SageMaker AI Studio-Ausführungsrolle an oder entfernen Sie sie, indem Sie die SAML Attribute analysieren.
- Passen Sie das Benutzerprofil an, indem Sie die Lebenszykluskonfiguration (LCC) hinzufügen und Tags hinzufügen.

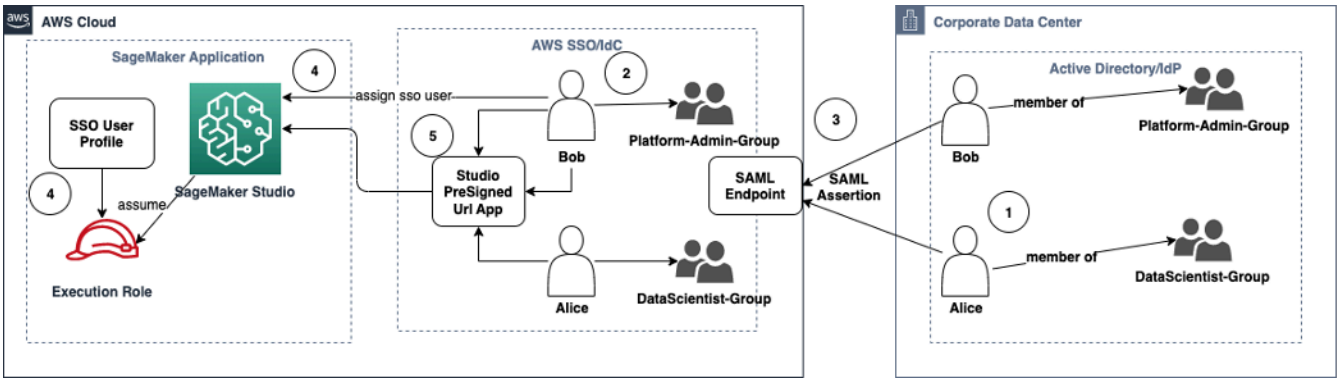
Zusammenfassend lässt sich sagen, dass diese Lösung SageMaker AI Studio als SAML2 2.0-Anwendung mit benutzerdefinierter Logik für Authentifizierung und Autorisierung verfügbar macht. Einzelheiten zur Implementierung finden Sie im Anhang, Abschnitt SageMaker Studio-Zugriff mithilfe von SAML Assertion.



Zugreifen auf SageMaker AI Studio mit einer benutzerdefinierten SAML Anwendung

AWSIAMIdC-Verbund

Die IdC-Verbundmethode ermöglicht es Kunden, sich von ihrem SAML IdP (wie Okta) direkt mit der SageMaker AI Studio-Anwendung zu verbinden. Das folgende Diagramm zeigt, wie der Verbundbenutzer autorisiert ist, auf seine eigene AI Studio-Instanz zuzugreifen. SageMaker



Zugriff auf SageMaker AI Studio im IdC-Modus IAM

1. Im Unternehmens-AD ist der Benutzer Mitglied von AD-Gruppen wie der Platform Admin-Gruppe und der Data Scientist-Gruppe.
2. Der AD-Benutzer und die AD-Gruppen von Identity Provider (IdP) werden mit AWS IAM Identity Center synchronisiert und sind als Single Sign-On-Benutzer bzw. Gruppen für Zuweisungen verfügbar.
3. Der IdP sendet eine SAML Assertion an den AWS SAML IdC-Endpunkt.
4. Im SageMaker AI Studio ist der iDC-Benutzer der Studio-Anwendung zugewiesen. SageMaker Diese Zuweisung kann mithilfe von IdC Group vorgenommen werden, und SageMaker AI Studio

gilt für jede iDC-Benutzerebene. Wenn diese Zuweisung erstellt wird, erstellt SageMaker AI Studio ein iDC-Benutzerprofil und weist die Rolle für die Domänenausführung zu.

5. Der Benutzer greift über die sichere, vorsignierte, als Cloud-Anwendung vom iDC URL gehostete sichere Anwendung auf die SageMaker AI Studio-Anwendung zu. SageMaker AI Studio übernimmt die Ausführungsrolle, die ihrem IdC-Benutzerprofil zugewiesen ist.

Anleitung zur Domänenauthentifizierung

Hier sind einige Überlegungen bei der Auswahl des Authentifizierungsmodus für eine Domain:

1. Wenn Sie möchten, dass Ihre Benutzer nicht direkt auf die SageMaker AI Studio-Benutzeroberfläche zugreifen AWS Management Console und diese aufrufen, verwenden Sie den Single Sign-On-Modus mit AWS IAM iDC.
2. Wenn Sie möchten, dass Ihre Benutzer nicht direkt im IAM Modus auf die SageMaker AI Studio-Benutzeroberfläche zugreifen AWS Management Console und diese anzeigen, können Sie dies tun, indem Sie eine Lambda-Funktion im Backend verwenden, um ein URL für das Benutzerprofil vorsigniertes Benutzerprofil zu generieren und sie zur AI Studio-Benutzeroberfläche weiterzuleiten. SageMaker
3. Im IdC-Modus wird jeder Benutzer einem einzelnen Benutzerprofil zugeordnet.
4. Allen Benutzerprofilen wird im IdC-Modus automatisch die Standard-Ausführungsrolle zugewiesen. Wenn Sie möchten, dass Ihren Benutzern unterschiedliche Ausführungsrollen zugewiesen werden, müssen Sie die Benutzerprofile mithilfe von aktualisieren. UpdateUserProfileAPI
5. Wenn Sie den Zugriff auf die SageMaker AI Studio-Benutzeroberfläche im IAM Modus (unter Verwendung des generierten, vorsigniertenURL) auf einen VPC Endpunkt beschränken möchten, ohne das Internet zu durchqueren, können Sie einen benutzerdefinierten Resolver verwenden. DNS Weitere Informationen finden Sie im Blogbeitrag Secure Amazon SageMaker AI Studio Presigned URLs Part 1: Fundamentale Infrastruktur.

Berechtigungsverwaltung

In diesem Abschnitt werden die bewährten Methoden für die Einrichtung häufig verwendeter IAM Rollen, Richtlinien und Leitplanken für die Bereitstellung und den Betrieb der AI Studio-Domain beschrieben. SageMaker

IAM-Rollen und -Richtlinien

Als bewährte Methode sollten Sie zunächst die relevanten Personen und Anwendungen, die sogenannten Principals, identifizieren, die am ML-Lebenszyklus beteiligt sind, und die AWS Berechtigungen, die Sie ihnen gewähren müssen. Da es sich bei SageMaker KI um einen verwalteten Dienst handelt, müssen Sie auch Service Principals berücksichtigen, d. AWS h. Dienste, die im Namen eines Benutzers API Anrufe tätigen können. Das folgende Diagramm zeigt die verschiedenen IAM Rollen, die Sie möglicherweise erstellen möchten, und zwar entsprechend den verschiedenen Personas in der Organisation.

SageMaker KI-Rollen IAM

Diese Rollen werden detailliert beschrieben, zusammen mit einigen spezifischen Beispielen, die IAMpermissions sie benötigen.

- ML-Admin-Benutzerrolle — Dies ist ein Principal, der die Umgebung für Datenwissenschaftler bereitstellt, indem er Studio-Domänen und Benutzerprofile (`sagemaker:CreateDomain`,`sagemaker:CreateUserProfile`) erstellt, Schlüssel AWS Key Management Service (AWS KMS) für Benutzer erstellt, S3-Buckets für Datenwissenschaftler erstellt und ECR Amazon-Repositorys für Container erstellt. Sie können auch Standardkonfigurationen und Lebenszyklusskripte für Benutzer festlegen, benutzerdefinierte Images erstellen und an die SageMaker AI Studio-Domain anhängen und Service Catalog-Produkte wie benutzerdefinierte Projekte und EMR Amazon-Vorlagen bereitstellen.

  Da dieser Schulleiter beispielsweise keine Schulungsaufträge ausführt, benötigt er keine Berechtigungen, um SageMaker KI-Schulungen oder Verarbeitungsjobs zu starten. Wenn sie

Infrastruktur als Codevorlagen verwenden, z. B. CloudFormation Terraform, um Domänen und Benutzer bereitzustellen, würde diese Rolle vom Bereitstellungsdienst übernommen, der die Ressourcen im Namen des Administrators erstellt. Diese Rolle hat möglicherweise nur Lesezugriff auf KI mithilfe von. SageMaker AWS Management Console

Diese Benutzerrolle benötigt außerdem bestimmte EC2 Berechtigungen, um die Domain in einem privaten Bereich zu startenVPC, KMS Berechtigungen zum Verschlüsseln des EFS Volumes sowie Berechtigungen zum Erstellen einer dienstverknüpften Rolle für Studio (). `iam:CreateServiceLinkedRole` Wir werden diese detaillierten Berechtigungen später in diesem Dokument beschreiben.

- Benutzerrolle „Data Scientist“ — Bei diesem Prinzip handelt es sich um den Benutzer, der sich bei SageMaker AI Studio anmeldet, die Daten untersucht, Verarbeitungs- und Schulungsaufträge und Pipelines erstellt usw. Die Hauptberechtigung, die der Benutzer benötigt, ist die Erlaubnis, SageMaker AI Studio zu starten. Die restlichen Richtlinien können von der SageMaker AI-Ausführungsdienstrolle verwaltet werden.
- SageMaker Rolle „KI-Ausführungsdienst“ — Da es sich bei SageMaker KI um einen verwalteten Dienst handelt, werden Jobs im Namen eines Benutzers gestartet. Diese Rolle ist häufig die umfassendste, was die erlaubten Berechtigungen angeht, da sich viele Kunden dafür entscheiden, eine einzige Ausführungsrolle zu verwenden, um Schulungsjobs, Verarbeitungsaufträge auszuführen oder Hosting-Jobs zu modellieren. Dies ist zwar ein einfacher Einstieg, da Kunden mit der Zeit reifer werden, aber sie teilen die Notebook-Ausführungsrolle häufig in separate Rollen für verschiedene API Aktionen auf, insbesondere wenn diese Jobs in bereitgestellten Umgebungen ausgeführt werden.

  Bei der Erstellung ordnen Sie der SageMaker AI Studio-Domäne eine Rolle zu. Da Kunden jedoch möglicherweise die Flexibilität benötigen, den verschiedenen Benutzerprofilen in der Domain unterschiedliche Rollen zuzuordnen (z. B. je nach ihrer beruflichen Funktion), können Sie jedem Benutzerprofil auch eine separate IAM Rolle zuordnen. Wir empfehlen, dass Sie einen einzelnen physischen Benutzer einem einzelnen Benutzerprofil zuordnen. Wenn Sie einem Benutzerprofil bei der Erstellung keine Rolle zuordnen, besteht das Standardverhalten darin, die Rolle für die SageMaker AIStudio Domänenausführung auch dem Benutzerprofil zuzuordnen.

  In Fällen, in denen mehrere Datenwissenschaftler und ML-Techniker gemeinsam an einem Projekt arbeiten und ein gemeinsames Berechtigungsmodell für den Zugriff auf Ressourcen benötigen, empfehlen wir Ihnen, eine SageMaker KI-Dienstausführungsrolle auf Teamebene zu erstellen, um die IAM Berechtigungen mit Ihren Teammitgliedern zu teilen. In den Fällen, in denen Sie die Berechtigungen auf jeder Benutzerebene sperren müssen, können Sie eine individuelle Rolle

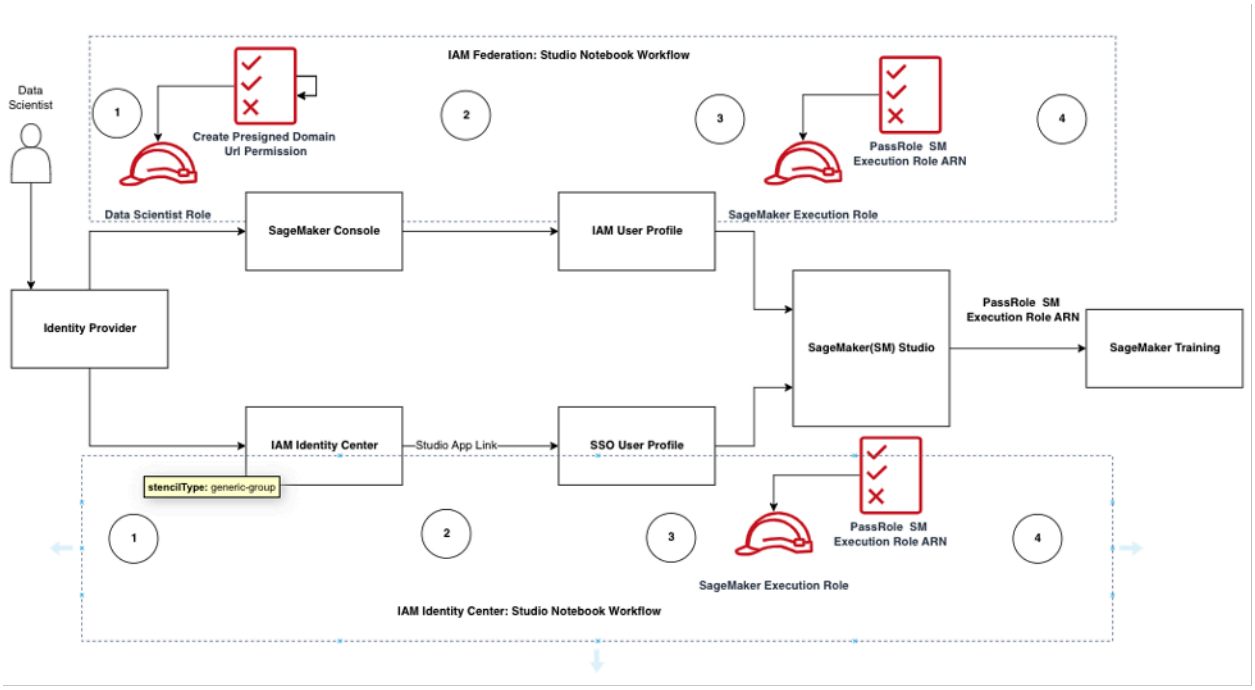
für die Ausführung von SageMaker KI-Diensten auf Benutzerebene erstellen. Dabei müssen Sie jedoch Ihre Dienstlimits beachten.

SageMaker Workflow für die Autorisierung von AI Studio Notebook

In diesem Abschnitt wird erläutert, wie die SageMaker AI Studio Notebook-Autorisierung für verschiedene Aktivitäten funktioniert, die der Data Scientist für die Erstellung und das Training des Modells direkt vom SageMaker AI Studio Notebook aus ausführen muss. Die SageMaker AI-Domäne unterstützt zwei Autorisierungsmodi:

- IAMFöderation
- IAMIdentitätszentrum

Als Nächstes führt Sie dieses paper durch den Data Scientist-Autorisierungsworkflow für jeden dieser Modi.



Der Authentifizierungs- und Autorisierungsablauf für Studio-Benutzer

IAMFederation: SageMaker Studio-Notebook-Workflows

1. Ein Data Scientist authentifiziert sich bei seinem Corporate Identity Provider und übernimmt die Data Scientist-Benutzerrolle (die Benutzerverbundrolle) in der SageMaker KI-Konsole. Diese

Verbundrolle hat die `iam:PassRole` API Berechtigung für die SageMaker AI-Ausführungsrolle, die Rolle Amazon Resource Name (ARN) an SageMaker Studio zu übergeben.

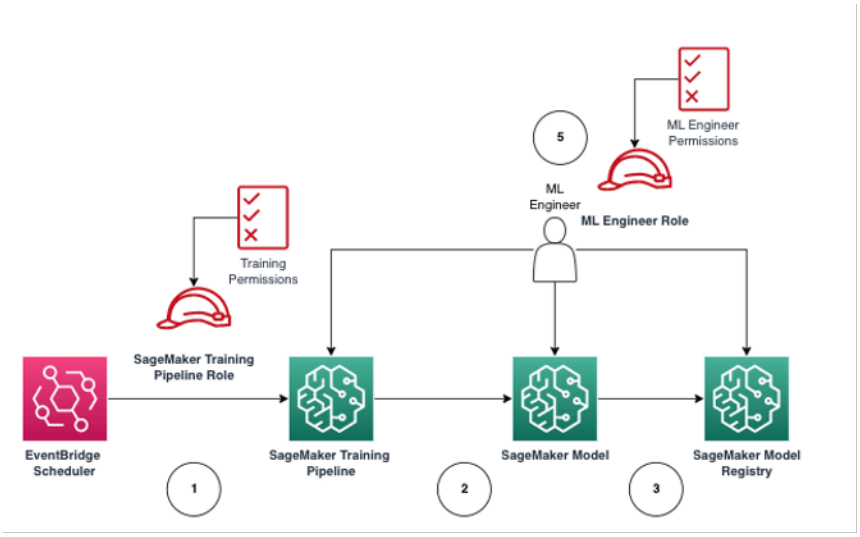
2. Der Data Scientist wählt den Open Studio-Link aus seinem IAM Studio-Benutzerprofil aus, der der SageMaker AI-Ausführungsrolle zugeordnet ist
3. Der SageMaker IDE Studio-Dienst wird gestartet, wobei die SageMaker Ausführungsrollenberechtigungen des Benutzerprofils vorausgesetzt werden. Diese Rolle hat für die SageMaker AI-Ausführungsrolle die `iam:PassRole` API Berechtigung, die Rolle ARN an den SageMaker KI-Schulungsdienst weiterzugeben.
4. Wenn Data Scientist den Trainingsjob in den Remote-Compute-Knoten startet, ARN wird die SageMaker KI-Ausführungsrolle an den SageMaker KI-Schulungsdienst übergeben. Dadurch wird eine neue Rollensitzung erstellt ARN und der Trainingsjob ausgeführt. Wenn Sie die Berechtigungen für einen Schulungsjob weiter einschränken müssen, können Sie eine spezielle Rolle für Schulungen erstellen und diese Rolle ARN beim Aufruf der Schulung weitergebenAPI.

IAMIdentity Center: Arbeitsablauf für Ihr Notizbuch mit SageMaker AI Studio

1. Der Data Scientist authentifiziert sich bei seinem Corporate Identity Provider und klickt auf AWS IAM Identity Center. Dem Datenwissenschaftler wird das Identity Center-Portal für den Benutzer angezeigt.
2. Der Data Scientist klickt auf den Link zur SageMaker AI Studio App, der in seinem IdC-Benutzerprofil erstellt wurde und der SageMaker AI-Ausführungsrolle zugeordnet ist.
3. Der SageMaker AI IDE Studio-Dienst wird gestartet, wobei die Berechtigungen des Benutzerprofils für die SageMaker AI-Ausführungsrolle vorausgesetzt werden. Diese Rolle hat für die SageMaker AI-Ausführungsrolle die `iam:PassRole` API Berechtigung, die Rolle ARN an den SageMaker KI-Schulungsdienst weiterzugeben.
4. Wenn der Data Scientist den Trainingsjob in Remote-Compute-Knoten startet, ARN wird die SageMaker KI-Ausführungsrolle an den SageMaker KI-Schulungsdienst übergeben. Die Ausführungsrolle ARN erstellt damit eine neue Rollensitzung und führt den Trainingsjob aus. ARN Wenn Sie die Berechtigungen für Trainingsjobs weiter einschränken müssen, können Sie eine schulungsspezifische Rolle erstellen und diese Rolle ARN beim Aufrufen der Schulung weitergeben. API

Bereitgestellte Umgebung: SageMaker KI-Schulungs-Workflow

In Bereitstellungsumgebungen wie Systemtests und Produktion werden Jobs über automatisierte Scheduler- und Event-Trigger ausgeführt, und der menschliche Zugriff auf diese Umgebungen ist von SageMaker AI Studio Notebooks aus eingeschränkt. In diesem Abschnitt wird erläutert, wie IAM Rollen mit der SageMaker KI-Trainingspipeline in der bereitgestellten Umgebung zusammenarbeiten.



SageMaker KI-Schulungsablauf in einer verwalteten Produktionsumgebung

1. Amazon EventBridge Scheduler löst den SageMaker KI-Trainingspipeline-Job aus.
2. Der SageMaker KI-Trainingspipeline-Job übernimmt die Rolle der SageMaker KI-Trainingspipeline, um das Modell zu trainieren.
3. Das trainierte SageMaker KI-Modell ist im SageMaker AI Model Registry registriert.
4. Ein ML-Ingenieur übernimmt die Benutzerrolle des ML-Ingenieurs, um die Trainingspipeline und das SageMaker KI-Modell zu verwalten.

Datenberechtigungen

Die Fähigkeit von SageMaker AI Studio-Benutzern, auf jede Datenquelle zuzugreifen, hängt von den Berechtigungen ab, die mit ihrer SageMaker IAM KI-Ausführungsrolle verknüpft sind. Die beigefügten Richtlinien können sie autorisieren, bestimmte Amazon S3 S3-Buckets oder -Präfixe zu lesen, zu schreiben oder zu löschen und eine Verbindung zu Amazon-Datenbanken herzustellen. RDS

Zugreifen auf Daten AWS Lake Formation

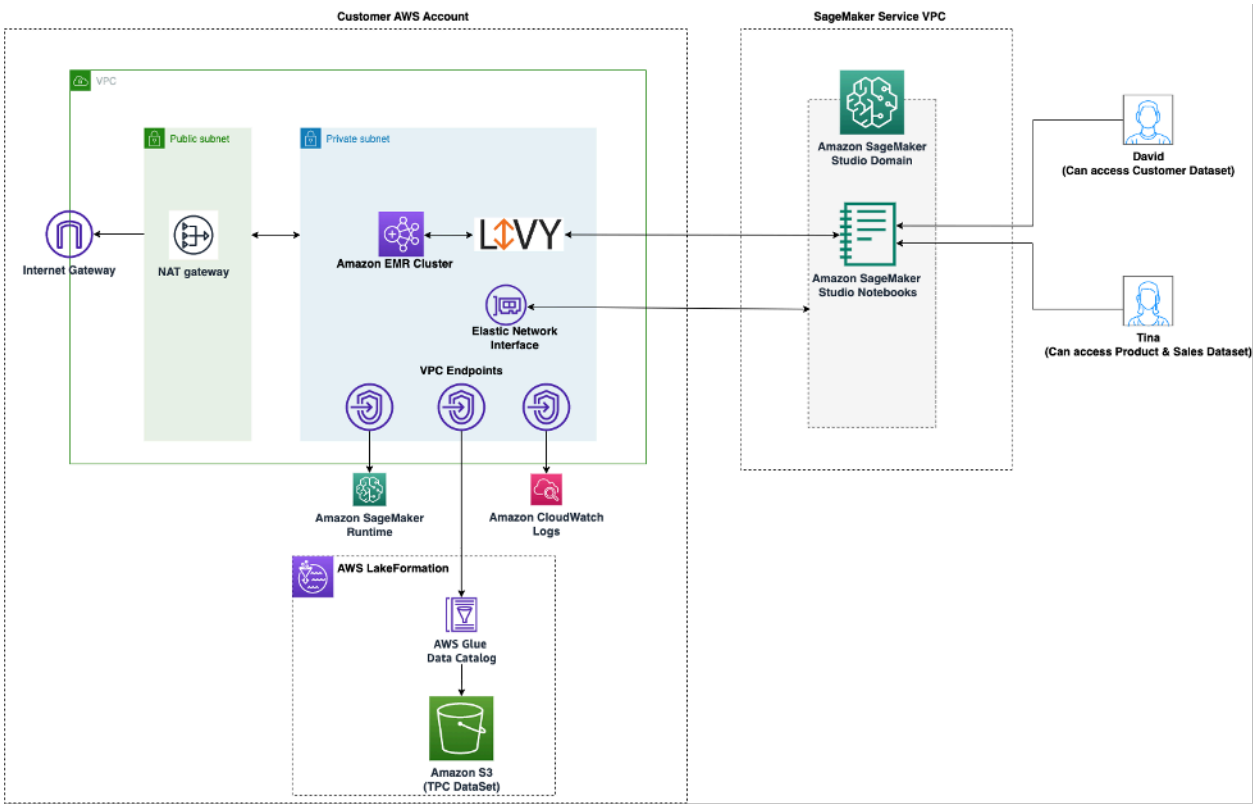
Viele Unternehmen haben damit begonnen, Data Lakes zu nutzen AWS Lake Formation, die von gesteuert werden, um ihren Benutzern einen detaillierten Datenzugriff zu ermöglichen. Als Beispiel für solche verwalteten Daten können Administratoren sensible Spalten für einige Benutzer maskieren und gleichzeitig Abfragen derselben Basistabelle ermöglichen.

Um Lake Formation von SageMaker AI Studio aus zu nutzen, können Administratoren SageMaker IAM KI-Ausführungsrollen als registrieren`DataLakePrincipals`. Weitere Informationen finden Sie unter Lake Formation Permissions Reference. Nach der Autorisierung gibt es drei Hauptmethoden für den Zugriff auf und das Schreiben von kontrollierten Daten aus SageMaker AI Studio:

1. Von einem SageMaker AI Studio Notebook aus können Benutzer Abfrage-Engines wie Amazon Athena oder Bibliotheken verwenden, die auf boto3 aufbauen, um Daten direkt auf das Notebook zu übertragen. Die AWSSDKfor Pandas (früher bekannt als awswrangler) ist eine beliebte Bibliothek. Im Folgenden finden Sie ein Codebeispiel, das zeigt, wie nahtlos dies sein kann:

   ```
   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. Verwenden Sie die native Konnektivität von SageMaker AI Studio zu Amazon, EMR um Daten in großem Umfang zu lesen und zu schreiben. Durch die Verwendung von Apache Livy- und EMR Amazon-Runtime-Rollen hat SageMaker AI Studio eine native Konnektivität aufgebaut, mit der Sie Ihre SageMaker IAM KI-Ausführungsrolle (oder eine andere autorisierte Rolle) für den Datenzugriff und die Datenverarbeitung an einen EMR Amazon-Cluster übergeben können. up-to-dateAnweisungen finden Sie unter Von Studio aus mit einem EMR Amazon-Cluster Connect.



Architektur für den Zugriff auf Daten, die von Lake Formation aus SageMaker Studio verwaltet werden

3. Verwenden Sie die native Konnektivität von SageMaker AI Studio für AWS Glue interaktive Sitzungen, um Daten in großem Umfang zu lesen und zu schreiben. SageMaker AI Studio Notebooks verfügen über integrierte Kernel, auf denen Benutzer Befehle interaktiv ausführen können. AWS Glue Dies ermöglicht die skalierbare Verwendung von Python-, Spark- oder Ray-Backends, mit denen Daten aus kontrollierten Datenquellen problemlos in großem Umfang gelesen und geschrieben werden können. Die Kernel ermöglichen es Benutzern, ihre SageMaker Ausführungs- oder andere autorisierte IAM Rollen zu übergeben. Weitere Informationen finden Sie unter Daten mithilfe AWS Glue interaktiver Sitzungen vorbereiten.

Gemeinsame Leitplanken

In diesem Abschnitt werden die am häufigsten verwendeten Leitplanken für die Steuerung Ihrer ML-Ressourcen mithilfe von IAM Richtlinien, Ressourcenrichtlinien, VPC Endpunktrichtlinien und Dienststeuerungsrichtlinien () beschrieben. SCPs

Beschränken Sie den Notebook-Zugriff auf bestimmte Instanzen

Diese Dienststeuerungsrichtlinie kann verwendet werden, um die Instanztypen einzuschränken, auf die Datenwissenschaftler bei der Erstellung von Studio-Notebooks Zugriff haben. Beachten Sie, dass jeder Benutzer die „System“ -Instanz benötigt, die berechtigt ist, die standardmäßige Jupyter Server-App zu erstellen, die AI Studio hostet SageMaker .

```
{
     "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"
                     ]
                 }
             }
         }
     ]
 }
```

Beschränken Sie nicht konforme AI Studio-Domänen SageMaker

Bei SageMaker AI Studio-Domains kann die folgende Dienststeuerungsrichtlinie verwendet werden, um den Datenverkehr für den Zugriff auf Kundenressourcen so zu erzwingen, dass dieser nicht über das öffentliche Internet, sondern über das VPC eines Kunden erfolgt:

```
{
     "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"
                }
            }
        }
    ]
 }
```

Beschränken Sie das Starten nicht autorisierter SageMaker KI-Images

Die folgende Richtlinie verhindert, dass ein Benutzer ein nicht autorisiertes SageMaker KI-Image in seiner Domain startet:f

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

Starten Sie Notebooks nur über KI-Endpunkte SageMaker VPC

Zusätzlich zu den VPC Endpunkten für die SageMaker KI-Kontrollebene unterstützt KI VPC Endpunkte, über die Benutzer eine Verbindung zu SageMaker AI Studio-Notebooks oder SageMaker SageMaker KI-Notebook-Instanzen herstellen können. Wenn Sie bereits einen VPC Endpunkt für eine SageMaker AI Studio-/Notebook-Instanz eingerichtet haben, erlaubt der folgende IAM Bedingungsschlüssel nur Verbindungen zu SageMaker AI Studio-Notebooks, wenn sie über den AI VPC Studio-Endpunkt oder über den SageMaker KI-Endpunkt hergestellt werden. SageMaker 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"
                     ]
                 }
             }
         }
     ]
 }
```

Beschränken Sie den Zugriff auf SageMaker AI Studio-Notebooks auf einen begrenzten IP-Bereich

Unternehmen beschränken den Zugriff auf SageMaker AI Studio häufig auf bestimmte zulässige Unternehmens-IP-Bereiche. Die folgende IAM Richtlinie mit dem `SourceIP` Bedingungsschlüssel kann dies einschränken.

```
{
```

```
    "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"
                    ]
                }
            }
        }
    ]
}
```

Verhindern Sie, dass SageMaker AI Studio-Benutzer auf andere Benutzerprofile zugreifen

Stellen Sie als Administrator bei der Erstellung des Benutzerprofils sicher, dass das Profil mit dem SageMaker AI Studio-Benutzernamen und dem Tag-Schlüssel gekennzeichnet ist`studiouserid`. Der Prinzipal (Benutzer oder Rolle, die dem Benutzer zugewiesen ist) sollte auch über ein Tag mit dem Schlüssel verfügen `studiouserid` (dieses Tag kann beliebig benannt werden und ist nicht darauf beschränkt`studiouserid`).

Fügen Sie als Nächstes der Rolle, die der Benutzer beim Start von SageMaker AI Studio annehmen wird, die folgende Richtlinie hinzu.

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

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

Tagging erzwingen

Datenwissenschaftler müssen SageMaker KI Studio-Notebooks verwenden, um Daten zu untersuchen und Modelle zu erstellen und zu trainieren. Das Anbringen von Tags auf Notebooks hilft bei der Überwachung der Nutzung und der Kostenkontrolle sowie bei der Sicherstellung der Eigentumsrechte und der Überprüfbarkeit.

Stellen Sie bei SageMaker AI Studio-Apps sicher, dass das Benutzerprofil markiert ist. Tags werden automatisch aus dem Benutzerprofil an Apps weitergegeben. Um die Erstellung von Benutzerprofilen mit Tags zu erzwingen (unterstützt durch CLI undSDK), sollten Sie erwägen, diese Richtlinie der Administratorrolle hinzuzufügen:

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

```
}
```

Für andere Ressourcen, wie z. B. Schulungsaufträge und Verarbeitungsaufträge, können Sie mithilfe der folgenden Richtlinie Kennzeichnungen zur Pflicht machen:

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

Root-Zugriff in SageMaker AI Studio

In SageMaker AI Studio läuft das Notebook in einem Docker-Container, der standardmäßig keinen Root-Zugriff auf die Host-Instanz hat. In ähnlicher Weise werden alle anderen Benutzer-ID-Bereiche innerhalb des Containers, mit Ausnahme des standardmäßigen Run-as-Users, auf der Host-Instanz selbst als Benutzer ohne Zugriffsrechte neu zugeordnet. IDs Daher beschränkt sich die Gefahr einer Rechteausweitung auf den Notebook-Container selbst.

Wenn Sie benutzerdefinierte Images erstellen, möchten Sie Ihrem Benutzer möglicherweise andere Rechte als Root-Rechte zur Verfügung stellen, um strengere Kontrollen zu ermöglichen. So können Sie beispielsweise verhindern, dass unerwünschte Prozesse als Root-Benutzer ausgeführt werden oder öffentlich verfügbare Pakete installiert werden. In solchen Fällen können Sie das Image so erstellen, dass es als Nicht-Root-Benutzer innerhalb der Docker-Datei ausgeführt wird. Unabhängig davon, ob Sie den Benutzer als Root oder als Nicht-Root erstellen, müssen Sie sicherstellen, dass

das UID/GID of the user is identical to the UID/GID AppImageConfigfür die benutzerdefinierte App gilt, die die Konfiguration für die SageMaker KI erstellt, um eine App mithilfe des benutzerdefinierten Images auszuführen. Wenn Ihr Dockerfile beispielsweise für einen Benutzer erstellt wurde, der kein Root-Benutzer ist, wie zum Beispiel den folgenden:

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

Die `AppImageConfig` Datei muss dasselbe erwähnen UID und GID darin: `KernelGatewayConfig`

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

Die akzeptablen GID Werte fürUID/für benutzerdefinierte Bilder sind 0/0 und 1000/100 für Studio-Bilder. Beispiele für die Erstellung benutzerdefinierter Images und die zugehörigen `AppImageConfig` Einstellungen finden Sie in diesem Github-Repository.

Um zu verhindern, dass Benutzer dies manipulieren, gewähren Sie Benutzern von SageMaker AI Studio-Notebooks keine `DeleteAppImageConfig` Berechtigungen wie `CreateAppImageConfigUpdateAppImageConfig`, oder.

Netzwerkmanagement

Um die SageMaker AI Studio-Domäne einzurichten, müssen Sie das VPC Netzwerk, die Subnetze und die Sicherheitsgruppen angeben. Achten Sie bei der Angabe der Subnetze VPC und darauf, dass Sie bei der Zuteilung das Nutzungsvolumen und das erwartete Wachstum IPs berücksichtigen, die in den folgenden Abschnitten erörtert werden.

VPCNetzwerkplanung

VPCKundensubnetze, die der SageMaker AI Studio-Domäne zugeordnet sind, müssen mit dem entsprechenden Classless Inter-Domain Routing (CIDR) -Bereich erstellt werden. Dies hängt von den folgenden Faktoren ab:

- Anzahl der Benutzer.
- Anzahl der Apps pro Benutzer.
- Anzahl der eindeutigen Instanztypen pro Benutzer.
- Durchschnittliche Anzahl von Trainingsinstanzen pro Benutzer.
- Erwartetes Wachstum in%

SageMaker KI und teilnehmende AWS Dienste fügen elastische Netzwerkschnittstellen (ENI) für die folgenden Anwendungsfälle in das VPC Kundensubnetz ein:

- Amazon EFS fügt ein ENI für ein EFS Mount-Ziel für die SageMaker AI-Domain ein (eine IP pro Subnetz/Availability Zone, die an die SageMaker AI-Domain angehängt ist).
- SageMaker AI Studio fügt ENI für jede einzelne Instanz, die von einem Benutzerprofil oder einem gemeinsam genutzten Bereich verwendet wird, eine ein. Beispielsweise:
  - Wenn ein Benutzerprofil eine standardmäßige Jupyter-Server-App (eine „System“ -Instanz), eine Data Science-App und eine Basis-Python-App (beide auf einer `ml.t3.medium` Instanz ausgeführt) ausführt, injiziert Studio zwei IP-Adressen.
  - Wenn ein Benutzerprofil eine standardmäßige Jupyter-Server-App (eine „System“ -Instanz), eine GPUTensorflow-App (auf einer `ml.g4dn.xlarge` Instanz) und eine Data Wrangler-App (auf einer `ml.m5.4xlarge` Instanz) ausführt, injiziert Studio drei IP-Adressen.
- ENIFür jeden VPC Endpunkt in VPC Domain-Subnetzen/Availability Zones wird eine eingefügt (vier IPs für SageMaker VPC KI-Endpunkte; IPs ~sechs für teilnehmende Dienstendpunkte wie S3, und.) VPC ECR CloudWatch

- Wenn SageMaker KI-Schulungs- und Verarbeitungsjobs mit derselben VPC Konfiguration gestartet werden, benötigt jeder Job zwei IP-Adressen pro Instanz.

> **Note**
>
> VPCEinstellungen für SageMaker AI Studio, wie Subnetze und VPC reiner Datenverkehr, werden nicht automatisch an die in AI Studio erstellten Schulungs- und Verarbeitungsaufträge weitergegeben. SageMaker Der Benutzer muss die VPC Einstellungen und die Netzwerkisolierung nach Bedarf einrichten, wenn er den APIs Create*Job aufruft. Weitere Informationen finden Sie unter Trainings- und Inferenzcontainer im internetfreien Modus ausführen.

Szenario: Data Scientist führt Experimente mit zwei verschiedenen Instance-Typen durch

Gehen Sie in diesem Szenario davon aus, dass eine SageMaker KI-Domäne im Modus „VPCNur Verkehr“ eingerichtet ist. Es sind VPC Endpunkte wie SageMaker KI, SageMaker AI RuntimeAPI, Amazon S3 und Amazon ECR eingerichtet.

Ein Datenwissenschaftler führt Experimente mit Studio-Notebooks durch, die auf zwei verschiedenen Instance-Typen (z. B. `ml.t3.medium` und`ml.m5.large`) ausgeführt werden, und startet zwei Apps in jedem Instance-Typ.

Nehmen wir an, der Datenwissenschaftler führt gleichzeitig einen Trainingsjob mit derselben VPC Konfiguration auf einer `ml.m5.4xlarge` Instanz aus.

In diesem Szenario wird der SageMaker AI Studio-Dienst ENIs wie folgt injiziert:

Tabelle 1 — Dem Kunden VPC für ein Experimentationsszenario ENIs eingespeist

Entität	Ziel	ENIeingespritzt	Hinweise	Level
EFSZiel montieren	VPCSubnetze	Drei	Drei AZs / Subnetze	Domain
VPC-Endpunkte	VPCSubnetze	30	Drei AZs / Subnetze mit jeweils 10 VPCE	Domain

Entität	Ziel	ENIeingespritzt	Hinweise	Level
Jupyter Server	VPC-Subnetz	One	Eine IP pro Instanz	Benutzer
KernelGateway App	VPC-Subnetz	Zwei	Eine IP pro Instanztyp	Benutzer
Training	VPC-Subnetz	Zwei	Zwei IPs pro Trainingsinstanz Fünf IPs pro Trainings instanz, falls EFAverwendet	Benutzer

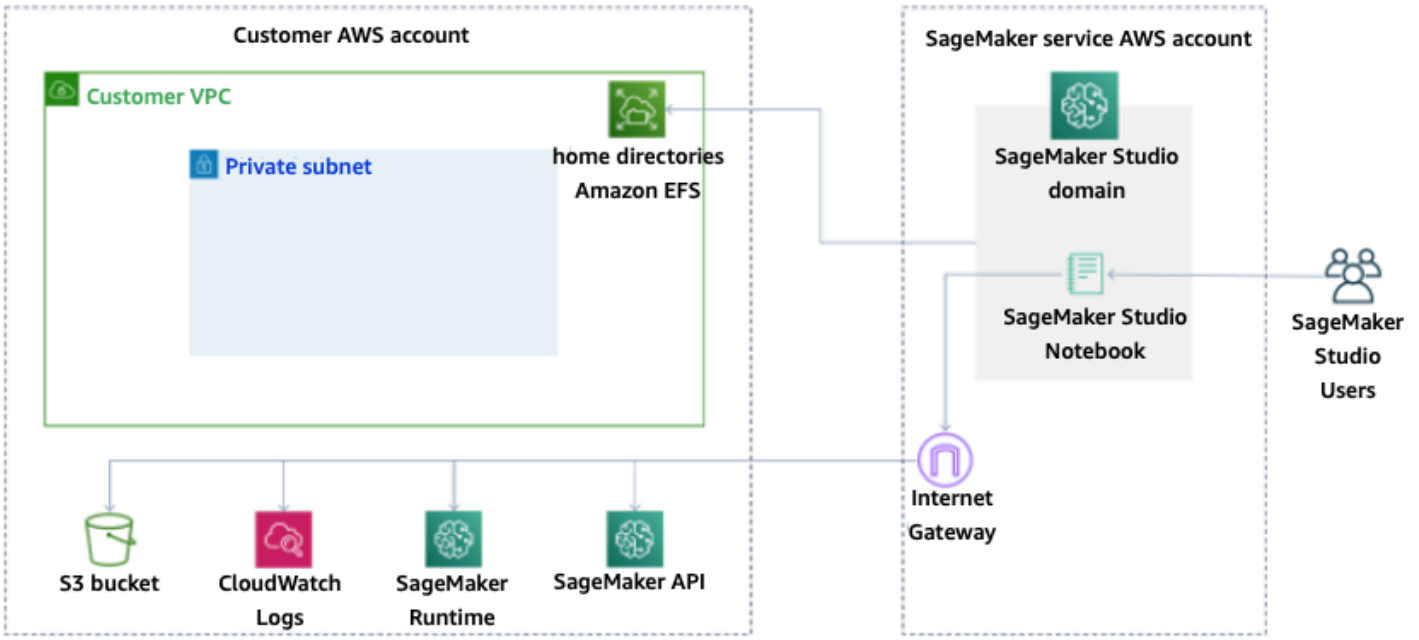
In diesem Szenario werden insgesamt 38 vom Kunden IPs konsumiert, VPC wobei 33 davon IPs von allen Benutzern auf Domänenebene und fünf auf Benutzerebene genutzt IPs werden. Wenn Sie 100 Benutzer mit ähnlichen Benutzerprofilen in dieser Domäne haben, die diese Aktivitäten gleichzeitig ausführen, verbrauchen Sie fünf x 100 = 500 IPs auf Benutzerebene, zusätzlich zum IP-Verbrauch auf Domänenebene, der 11 IPs pro Subnetz beträgt, also insgesamt 511. IPs Für dieses Szenario müssen Sie das VPC Subnetz CIDR mit /22 erstellen, das 1024 IP-Adressen zuweist, sodass weitere Optionen zur Verfügung stehen.

VPCNetzwerkoptionen

Eine SageMaker AI Studio-Domain unterstützt die Konfiguration des VPC Netzwerks mit einer der folgenden Optionen:

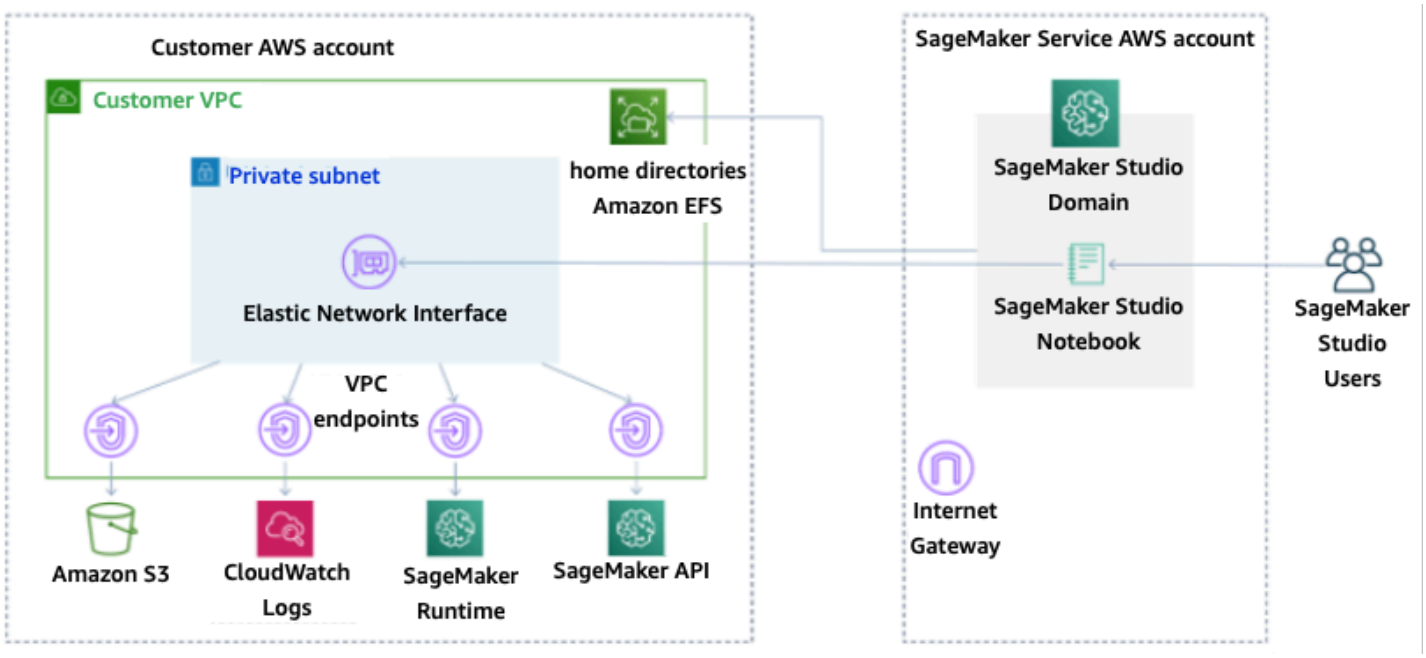
- Nur öffentliches Internet
- Nur VPC

Mit der Option „Nur öffentliches Internet“ können SageMaker API KI-Dienste das öffentliche Internet über das Internet-Gateway nutzenVPC, das im vom SageMaker AI-Dienstkonto verwalteten Internet-Gateway bereitgestellt wird, wie in der folgenden Abbildung dargestellt:



Standardmodus: Internetzugang über ein SageMaker AI-Servicekonto

Die VPCeinzige Option deaktiviert das Internet-Routing über das von der SageMaker KI VPC verwaltete Dienstkonto und ermöglicht es dem Kunden, den Datenverkehr so zu konfigurieren, dass er über VPC Endpunkte geleitet wird, wie in der folgenden Abbildung dargestellt:



VPCeinziger Modus: Kein Internetzugang über ein SageMaker AI-Servicekonto

Richten Sie für eine Domain, die im Modus „VPCNur“ eingerichtet ist, eine Sicherheitsgruppe pro Benutzerprofil ein, um eine vollständige Isolierung der zugrunde liegenden Instanzen zu gewährleisten. Jede Domäne in einem AWS Konto kann ihre eigene VPC Konfiguration und ihren eigenen Internetmodus haben. Weitere Informationen zur Einrichtung der VPC Netzwerkkonfiguration finden Sie unter Connect von SageMaker AI Studio-Notebooks in a VPC mit externen Ressourcen.

Einschränkungen

- Nachdem eine SageMaker AI Studio-Domäne erstellt wurde, können Sie der Domäne keine neuen Subnetze zuordnen.
- Der VPC Netzwerktyp (nur öffentliches Internet oder VPC nur) kann nicht geändert werden.

Datenschutz

Bevor ein ML-Workload konzipiert wird, sollten die grundlegenden Verfahren, die die Sicherheit beeinflussen, vorhanden sein. Die Datenklassifizierung bietet beispielsweise die Möglichkeit, Daten anhand ihrer Vertraulichkeitsstufen zu kategorisieren, und Verschlüsselung schützt Daten, indem sie sie für unbefugten Zugriff unverständlich macht. Diese Methoden sind wichtig, weil sie Ziele wie die Verhinderung von Missbrauch oder die Einhaltung gesetzlicher Verpflichtungen unterstützen.

SageMaker AI Studio bietet mehrere Funktionen zum Schutz von Daten im Speicher und bei der Übertragung. Wie im Modell der AWS gemeinsamen Verantwortung beschrieben, sind Kunden jedoch dafür verantwortlich, die Kontrolle über die Inhalte zu behalten, die auf der AWS globalen Infrastruktur gehostet werden. In diesem Abschnitt beschreiben wir, wie Kunden diese Funktionen zum Schutz ihrer Daten nutzen können.

Schützen Sie Daten im Ruhezustand

Um Ihre SageMaker AI Studio-Notebooks sowie Ihre Modellerstellungsdaten und Modellartefakte zu schützen, verschlüsselt SageMaker KI die Notizbücher sowie die Ergebnisse von Trainings- und Batch-Transformationsjobs. SageMaker AI verschlüsselt diese standardmäßig mit dem AWS Managed Key für Amazon S3. Dieser AWS verwaltete Schlüssel für Amazon S3 kann nicht für den kontoübergreifenden Zugriff freigegeben werden. Geben Sie für den kontoübergreifenden Zugriff Ihren vom Kunden verwalteten Schlüssel bei der Erstellung von SageMaker KI-Ressourcen an, damit er für den kontoübergreifenden Zugriff gemeinsam genutzt werden kann.

Mit SageMaker AI Studio können Daten an den folgenden Orten gespeichert werden:

- S3-Bucket — Wenn ein gemeinsam nutzbares Notizbuch aktiviert ist, teilt SageMaker AI Studio Notebook-Snapshots und Metadaten in einem S3-Bucket.
- EFSVolumen — SageMaker AI Studio fügt Ihrer Domain ein EFS Volume zum Speichern von Notizbüchern und Datendateien hinzu. Dieses EFS Volumen bleibt auch nach dem Löschen der Domain bestehen.
- EBSVolume — EBS ist an die Instanz angehängt, auf der das Notebook läuft. Dieses Volume bleibt für die Dauer der Instanz bestehen.

Verschlüsselung im Ruhezustand mit AWS KMS

- Sie können Ihren AWS KMS Schlüssel weitergeben, um ein EBS Volume zu verschlüsseln, das an Notebooks, Schulungen, Tuning, Batch-Transformationsaufträge und Endgeräte angeschlossen ist.
- Wenn Sie keinen KMS Schlüssel angeben, verschlüsselt SageMaker KI sowohl Betriebssystemvolumes (OS) als auch ML-Datenvolumes mit einem vom System verwalteten Schlüssel. KMS
- Vertrauliche Daten, die aus Compliance-Gründen mit einem KMS Schlüssel verschlüsselt werden müssen, sollten auf dem ML-Speichervolume oder in Amazon S3 gespeichert werden. Beide können mit einem von Ihnen angegebenen KMS Schlüssel verschlüsselt werden.

Schutz der Daten während der Übertragung

SageMaker AI Studio stellt sicher, dass ML-Modellartefakte und andere Systemartefakte bei der Übertragung und im Speicher verschlüsselt werden. Anfragen an die SageMaker KI API und die Konsole werden über eine sichere (SSL) Verbindung gestellt. Einige Daten innerhalb des Netzwerks sind während der Übertragung (innerhalb der Service-Plattform) unverschlüsselt. Dies umfasst:

- Kommunikation zwischen der Service-Steuerebene und Trainingsauftrags-Instances (keine Kundendaten).
- Kommunikation zwischen Knoten bei verteilten Verarbeitungs- und Trainingsaufgaben (netzwerkintern).

Sie können sich jedoch dafür entscheiden, die Kommunikation zwischen Knoten in einem Trainingscluster zu verschlüsseln. Die Verschlüsselung des Datenverkehrs zwischen Containern zu aktivieren, kann die Trainingszeit erhöhen, vor allem wenn Sie mit verteilten Deep-Learning-Algorithmen arbeiten.

Standardmäßig führt Amazon SageMaker AI Trainingsjobs in einem Amazon durchVPC, um die Sicherheit Ihrer Daten zu gewährleisten. Sie können eine weitere Sicherheitsstufe hinzufügen, um Ihre Trainingscontainer und -daten zu schützen, indem Sie eine private Sicherheitsstufe konfigurierenVPC. Darüber hinaus können Sie Ihre SageMaker AI Studio-Domain so konfigurieren, dass sie VPC nur im Modus ausgeführt wird, und VPC Endpunkte einrichten, um den Datenverkehr über ein privates Netzwerk weiterzuleiten, ohne dass ausgehender Datenverkehr über das Internet übertragen wird.

Leitplanken zum Datenschutz

Verschlüsseln Sie SageMaker KI-Hosting-Volumes im Ruhezustand

Verwenden Sie die folgende Richtlinie, um die Verschlüsselung beim Hosten eines SageMaker KI-Endpunkts für Online-Inferenz durchzusetzen:

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

Verschlüsseln Sie S3-Buckets, die während der Modellüberwachung verwendet werden

Model Monitoring erfasst Daten, die an Ihren SageMaker KI-Endpunkt gesendet werden, und speichert sie in einem S3-Bucket. Wenn Sie die Data Capture Config einrichten, müssen Sie den S3-Bucket verschlüsseln. Derzeit gibt es dafür keine kompensierende Kontrolle.

Der Model Monitoring-Service erfasst nicht nur die Ergebnisse der Endgeräte, sondern prüft auch, ob Abweichungen von einem vorher festgelegten Ausgangswert vorliegen. Sie müssen die Ausgaben und die Zwischenspeichervolumes, die zur Überwachung der Abweichung verwendet werden, verschlüsseln.

```
{
```

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

Verschlüsseln Sie ein SageMaker AI Studio-Domain-Speichervolume

Erzwingen Sie die Verschlüsselung des Speichervolumes, das an die Studio-Domain angehängt ist. Gemäß dieser Richtlinie muss ein Benutzer eine angeben, CMK um die an Studio-Domänen angehängten Speichervolumes zu verschlüsseln.

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

```
}
```

Verschlüsseln Sie in S3 gespeicherte Daten, die zur gemeinsamen Nutzung von Notizbüchern verwendet werden

Dies ist die Richtlinie zur Verschlüsselung aller im Bucket gespeicherten Daten, die für die gemeinsame Nutzung von Notizbüchern zwischen Benutzern in einer SageMaker AI Studio-Domain verwendet werden:

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

Einschränkungen

- Sobald eine Domain erstellt wurde, können Sie den angehängten EFS Volume-Speicher nicht mehr mit einem benutzerdefinierten AWS KMS Schlüssel aktualisieren.
- Sie können Trainings-/Verarbeitungsjobs oder Endpunktkonfigurationen nicht mit KMS Schlüsseln aktualisieren, nachdem sie einmal erstellt wurden.

Protokollierung und Überwachung

Um Ihnen beim Debuggen Ihrer Kompilierungs-, Verarbeitungs-, Trainingsjobs, Endpunkte, Transformationsjobs, Notebook-Instances und Lebenszykluskonfigurationen für Notebook-Instances zu helfen, wird alles, was ein Algorithmuscontainer, ein Modellcontainer oder eine Notebook-Instance-Lebenszykluskonfiguration an stdout oder stderr sendet, auch an Amazon Logs gesendet. CloudWatch Sie können SageMaker AI Studio mit Amazon überwachen CloudWatch, das Rohdaten sammelt und sie zu lesbaren Metriken verarbeitet, die nahezu in Echtzeit verfügbar sind. Diese Statistiken werden 15 Monate lang aufbewahrt, sodass Sie auf historische Informationen zugreifen und sich einen besseren Überblick über die Leistung Ihrer Webanwendung oder Ihres Dienstes verschaffen können.

Protokollierung mit CloudWatch

Da der datenwissenschaftliche Prozess von Natur aus experimentell und iterativ ist, ist es wichtig, Aktivitäten wie die Notebook-Nutzung, die Laufzeit von Schulungs- und Verarbeitungsjobs, Trainingsmetriken und Messwerte für die Endpunktbereitstellung wie die Aufruflatenz zu protokollieren. Standardmäßig veröffentlicht SageMaker KI Metriken in CloudWatch Logs, und diese Protokolle können mithilfe von vom Kunden verwalteten Schlüsseln verschlüsselt werden. AWS KMS

Sie können auch VPC Endpunkte verwenden, um Protokolle zu senden, CloudWatch ohne das öffentliche Internet zu nutzen. Sie können auch Alarme einrichten, die auf bestimmte Grenzwerte achten und Benachrichtigungen senden oder Aktivitäten auslösen, wenn diese Grenzwerte erreicht werden. Weitere Informationen finden Sie im CloudWatch Amazon-Benutzerhandbuch.

SageMaker AI erstellt eine einzelne Protokollgruppe für Studio, unter`/aws/sagemaker/studio`. Jedes Benutzerprofil und jede App hat ihren eigenen Protokollstream unter dieser Protokollgruppe, und auch Skripts zur Lebenszykluskonfiguration haben ihren eigenen Protokollstream. Ein Benutzerprofil mit dem Namen „studio-user“ mit einer Jupyter Server-App und einem angehängten Lifecycle-Skript und einer Data Science Kernel Gateway-App enthält beispielsweise die folgenden Protokollstreams:

```
/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
```

Damit SageMaker KI in Ihrem Namen Logs CloudWatch an Sie senden kann, benötigt der Aufrufer des Training/Processing/Transform Jobs APIs die folgenden Berechtigungen:

```
{
     "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"
         }
     ]
 }
```

Um diese Protokolle mit einem benutzerdefinierten AWS KMS Schlüssel zu verschlüsseln, müssen Sie zunächst die Schlüsselrichtlinie so ändern, dass der CloudWatch Dienst den Schlüssel ver- und entschlüsseln kann. Nachdem Sie einen AWS KMS Schlüssel zur Protokollverschlüsselung erstellt haben, ändern Sie die Schlüsselrichtlinie so, dass sie Folgendes umfasst:

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

Beachten Sie, dass Sie jederzeit einen bestimmten Amazon-Ressourcennamen (ARN) für das CloudWatch Protokoll, das Sie verschlüsseln möchten, verwenden `ArnEquals` und angeben können. Hier zeigen wir der Einfachheit halber, dass Sie diesen Schlüssel verwenden können, um alle Protokolle in einem Konto zu verschlüsseln. Darüber hinaus veröffentlichen Trainings-, Verarbeitungs- und Modellendpunkte Metriken über die Instanz CPU - und Speicherauslastung, die Latenz bei Hosting-Aufrufen usw. Sie können Amazon außerdem so konfigurierenSNS, dass Administratoren über Ereignisse informiert werden, wenn bestimmte Schwellenwerte überschritten werden. Der Nutzer der Schulung und Verarbeitung APIs muss über die folgenden Berechtigungen verfügen:

```
{
     "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"
        }
    ]
}
```

Prüfung mit AWS CloudTrail

Um Ihre Einhaltung von Vorschriften zu verbessern, sollten Sie alle APIs mit prüfen AWS CloudTrail. Standardmäßig APIs werden alle SageMaker KI protokolliert AWS CloudTrail. Für die Aktivierung benötigen Sie keine zusätzlichen IAM Berechtigungen CloudTrail.

Alle SageMaker KI-Aktionen, mit Ausnahme von `InvokeEndpoint` und`InvokeEndpointAsync`, werden von den Vorgängen protokolliert CloudTrail und in diesen dokumentiert. Beispielsweise generieren Aufrufe der `CreateNotebookInstance` Aktionen`CreateTrainingJob`,`CreateEndpoint`, und Einträge in den CloudTrail Protokolldateien.

Jeder CloudTrail Ereigniseintrag enthält Informationen darüber, wer die Anfrage generiert hat. Die Identitätsinformationen unterstützen Sie bei der Ermittlung der folgenden Punkte:

- Ob die Anforderung mit Root- oder AWS IAM-Benutzeranmeldeinformationen ausgeführt wurde.
- Gibt an, ob die Anforderung mit temporären Sicherheitsanmeldeinformationen für eine Rolle oder einen Verbundbenutzer gesendet wurde.
- Ob die Anfrage von einem anderen AWS Dienst gestellt wurde. Ein Beispielereignis finden Sie in der CloudTrail Dokumentation Log SageMaker AI API Calls.

Standardmäßig wird der Name der Studio-Ausführungsrolle des Benutzerprofils als ID für jedes Ereignis CloudTrail protokolliert. Dies funktioniert, wenn jeder Benutzer seine eigene Ausführungsrolle hat. Wenn sich mehrere Benutzer dieselbe Ausführungsrolle teilen, können Sie die `sourceIdentity` Konfiguration verwenden, um den Namen des Studio-Benutzerprofils weiterzugeben CloudTrail. Informationen zur Aktivierung der `sourceIdentity` Funktion finden Sie unter Überwachen des Zugriffs auf Benutzerressourcen von Amazon SageMaker AI Studio aus. In einem gemeinsam genutzten Bereich beziehen sich alle Aktionen auf den Bereich ARN als Quelle, und Sie können ihn nicht überprüfen`sourceIdentity`.

Kostenzuweisung

SageMaker AI Studio verfügt über integrierte Funktionen, mit denen Administratoren die Ausgaben ihrer einzelnen Domains, gemeinsam genutzten Bereiche und Benutzer verfolgen können.

Automatisiertes Tagging

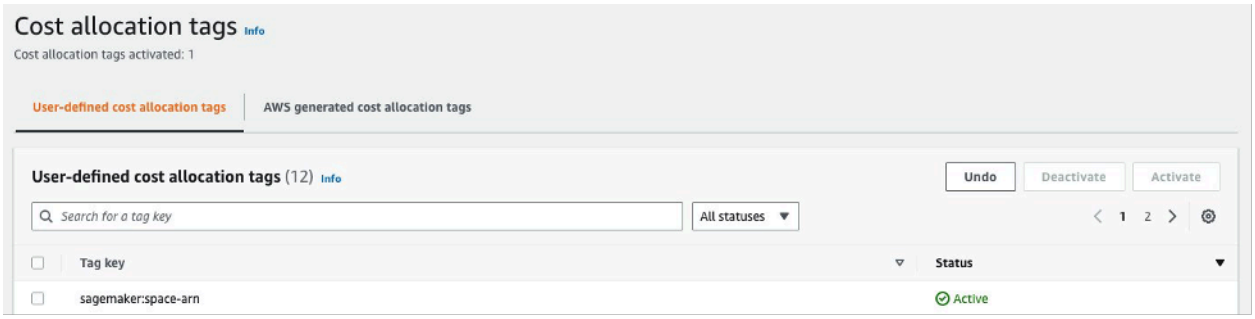
SageMaker AI Studio kennzeichnet neue SageMaker Ressourcen wie Trainingsjobs, Verarbeitungsjobs und Kernel-Apps jetzt automatisch mit den jeweiligen `sagemaker:domain-arn` Ressourcen. Auf einer detaillierteren Ebene kennzeichnet SageMaker KI die Ressource auch mit dem `sagemaker:user-profile-arn` Oder, `sagemaker:space-arn` um den Hauptersteller der Ressource zu bestimmen.

SageMaker EFSAI-Domänen-Volumes werden mit einem Schlüssel gekennzeichnet, der nach dem Wert der Domain benannt `ManagedByAmazonSageMakerResource` ist. ARN Sie verfügen nicht über detaillierte Tags, mit denen sich die Speicherplatznutzung auf Benutzerebene nachvollziehen lässt. Administratoren können das EFS Volume jedoch zur maßgeschneiderten Überwachung an eine EC2 Instanz anhängen.

Kostenüberwachung

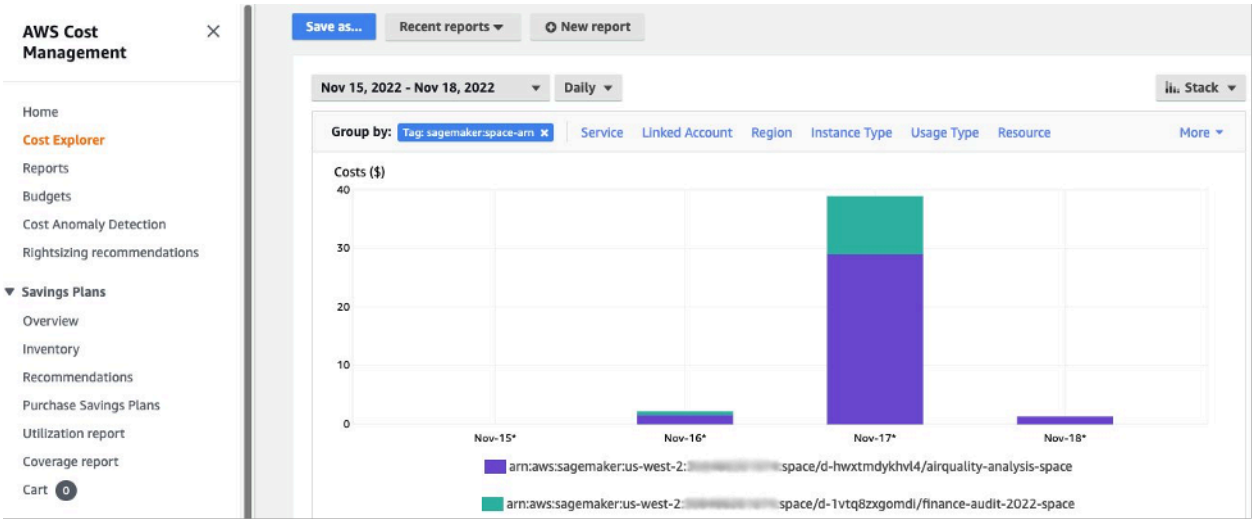
Automatisierte Tags ermöglichen es Administratoren, Ihre ML-Ausgaben mithilfe von out-of-the-box Lösungen wie und sowie benutzerdefinierten Lösungen AWS Budgets, die auf den Daten aus AWS Kosten AWS Cost Explorer- und Nutzungsberichten (CURs) basieren, nachzuverfolgen, zu melden und zu überwachen.

Um die angehängten Tags für die Kostenanalyse verwenden zu können, müssen sie zunächst in der AWS Billing Konsole im Bereich Kostenzuweisungs-Tags aktiviert werden. Es kann bis zu 24 Stunden dauern, bis Tags im Bereich „Kostenzuweisungstag“ angezeigt werden. Sie müssen also eine SageMaker KI-Ressource erstellen, bevor Sie sie aktivieren können.



Als Kostenzuordnungs-Tags im Cost Explorer ARN aktivierter Speicherplatz

Nachdem Sie ein Kostenzuordnungs-Tag aktiviert haben, AWS beginnt die Erfassung Ihrer markierten Ressourcen. Nach 24-48 Stunden werden die Tags im Kosten-Explorer als auswählbare Filter angezeigt.



Kosten gruppiert nach gemeinsam genutztem Speicherplatz für eine Beispieldomäne

Kostenkontrolle

Wenn der erste SageMaker AI Studio-Benutzer eingebunden ist, erstellt SageMaker AI ein EFS Volume für die Domain. Für dieses EFS Volumen fallen Speicherkosten an, da Notizbücher und Datendateien im Home-Verzeichnis des Benutzers gespeichert werden. Wenn der Benutzer Studio-Notebooks startet, werden sie für die Recheninstanzen gestartet, auf denen die Notebooks ausgeführt werden. Eine detaillierte Aufschlüsselung der Kosten finden Sie in der Amazon SageMaker AI-Preisübersicht.

Administratoren können die Rechenkosten kontrollieren, indem sie die Liste der Instances angeben, die ein Benutzer einrichten kann. Dabei verwenden sie die IAM Richtlinien, die im Abschnitt Allgemeine Richtlinien beschrieben sind. Darüber hinaus empfehlen wir Kunden, die SageMaker AI Studio-Erweiterung zum auto Herunterfahren zu verwenden, um Kosten zu sparen, indem inaktive Apps automatisch heruntergefahren werden. Diese Servererweiterung fragt regelmäßig pro Benutzerprofil nach laufenden Apps ab und fährt inaktive Apps auf der Grundlage eines vom Administrator festgelegten Timeouts herunter.

Um diese Erweiterung für alle Benutzer in Ihrer Domain festzulegen, können Sie eine Lebenszykluskonfiguration verwenden, wie im Abschnitt Anpassung beschrieben. Darüber hinaus

können Sie auch den Extension Checker verwenden, um sicherzustellen, dass alle Benutzer Ihrer Domain die Erweiterung installiert haben.

Anpassung

Lebenszyklus-Konfiguration

Lebenszykluskonfigurationen sind Shell-Skripte, die durch SageMaker AI Studio-Lebenszyklusereignisse initiiert werden, z. B. durch das Starten eines neuen SageMaker AI Studio-Notebooks. Sie können diese Shell-Skripts verwenden, um die Anpassung für Ihre SageMaker AI Studio-Umgebungen zu automatisieren, z. B. die Installation benutzerdefinierter Pakete, die Jupyter-Erweiterung für das automatische Herunterfahren inaktiver Notebook-Apps und die Einrichtung der Git-Konfiguration. Detaillierte Anweisungen zum Erstellen von Lebenszykluskonfigurationen finden Sie in diesem Blog: Anpassen von Amazon SageMaker AI Studio mithilfe von Lebenszykluskonfigurationen.

Benutzerdefinierte Bilder für SageMaker AI Studio-Notebooks

Studio-Notebooks werden mit einer Reihe von vorgefertigten Images geliefert, die aus Amazon SageMaker AI Python SDK und der neuesten Version der IPython Runtime oder des Kernels bestehen. Mit dieser Funktion können Sie Ihre eigenen benutzerdefinierten Bilder auf Amazon SageMaker AI-Notizbücher übertragen. Diese Bilder stehen dann allen Benutzern zur Verfügung, die in der Domain authentifiziert sind.

Entwickler und Datenwissenschaftler benötigen möglicherweise benutzerdefinierte Images für verschiedene Anwendungsfälle:

- Zugriff auf bestimmte oder aktuelle Versionen beliebter ML-Frameworks wie TensorFlowMXNet, PyTorch, oder andere.
- Bringen Sie benutzerdefinierten Code oder lokal entwickelte Algorithmen in SageMaker AI Studio-Notebooks für schnelle Iterationen und Modelltraining.
- Zugriff auf Data Lakes oder lokale Datenspeicher über. APIs Administratoren müssen die entsprechenden Treiber in das Image aufnehmen.
- Zugriff auf eine andere Backend-Laufzeit (auch Kernel genannt) als IPython (wie R, Julia oder andere). Sie können auch den beschriebenen Ansatz verwenden, um einen benutzerdefinierten Kernel zu installieren.

Detaillierte Anweisungen zum Erstellen eines benutzerdefinierten Images finden Sie unter Benutzerdefiniertes SageMaker AI-Image erstellen.

JupyterLab Erweiterungen

Mit SageMaker AI Studio JuypterLab 3 Notebook können Sie die Vorteile der ständig wachsenden Community von JupyterLab Open-Source-Erweiterungen nutzen. In diesem Abschnitt werden einige vorgestellt, die sich auf natürliche Weise in den SageMaker KI-Entwickler-Workflow einfügen. Wir empfehlen Ihnen jedoch, die verfügbaren Erweiterungen zu durchsuchen oder sogar Ihre eigenen zu erstellen.

JupyterLab 3 macht das Paketieren und Installieren von Erweiterungen jetzt erheblich einfacher. Sie können die oben genannten Erweiterungen über Bash-Skripte installieren. Öffnen Sie beispielsweise in SageMaker AI Studio das Systemterminal über den Studio-Launcher und führen Sie die folgenden Befehle aus. Darüber hinaus können Sie die Installation dieser Erweiterungen mithilfe von Lebenszykluskonfigurationen automatisieren, sodass sie zwischen den Neustarts von Studio beibehalten werden. Sie können dies für alle Benutzer in der Domäne oder auf individueller Benutzerebene konfigurieren.

Um beispielsweise eine Erweiterung für einen Amazon S3 S3-Dateibrowser zu installieren, führen Sie die folgenden Befehle im Systemterminal aus und stellen Sie sicher, dass Sie Ihren Browser aktualisieren:

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

Weitere Informationen zur Erweiterungsverwaltung, einschließlich der Erstellung von Lebenszykluskonfigurationen, die aus Gründen der Abwärtskompatibilität sowohl für die Versionen 1 als auch für 3 von JupyterLab Notebooks funktionieren, finden Sie unter Installation JupyterLab und Jupyter Server-Erweiterungen.

Git-Repositorien

SageMaker AI Studio ist mit einer Jupyter-Git-Erweiterung vorinstalliert, mit der Benutzer ein maßgeschneidertes URL Git-Repository aufrufen, es in Ihr EFS Verzeichnis klonen, Änderungen übertragen und den Commit-Verlauf anzeigen können. Administratoren können vorgeschlagene

Git-Repos auf Domainebene so konfigurieren, dass sie den Endbenutzern als Drop-down-Optionen angezeigt werden. up-to-dateAnweisungen finden Sie unter Vorgeschlagene Git-Repos an Studio anhängen.

Wenn ein Repository privat ist, fordert die Erweiterung den Benutzer auf, seine Anmeldeinformationen mithilfe der Standard-Git-Installation in das Terminal einzugeben. Alternativ kann der Benutzer zur einfacheren Verwaltung SSH-Anmeldeinformationen in seinem individuellen EFS Verzeichnis speichern.

Conda-Umgebung

SageMaker AI Studio-Notebooks verwenden Amazon EFS als persistente Speicherschicht. Datenwissenschaftler können den persistenten Speicher nutzen, um benutzerdefinierte Conda-Umgebungen zu erstellen und diese Umgebungen zur Erstellung von Kernels zu verwenden. Diese Kernel werden von Kernel-EFS, App- oder Studio-Neustarts unterstützt und bleiben auch zwischen Kerneln-, App- oder Studio-Neustarts bestehen. Studio nimmt automatisch alle gültigen Umgebungen als KernelGateway Kernel auf.

Das Erstellen einer Conda-Umgebung ist für einen Datenwissenschaftler unkompliziert, aber es dauert etwa eine Minute, bis die Kernel im Kernel-Selektor aufgefüllt sind. Um eine Umgebung zu erstellen, führen Sie den folgenden Befehl in einem Systemterminal aus:

```
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
```

Eine ausführliche Anleitung finden Sie im Abschnitt Perist Conda environments to the Studio EFS volume unter Vier Ansätze zur Verwaltung von Python-Paketen in Amazon SageMaker Studio-Notebooks.

Schlussfolgerung

In diesem Whitepaper haben wir verschiedene Best Practices in Bereichen wie Betriebsmodell, Domainmanagement, Identitätsmanagement, Berechtigungsmanagement, Netzwerkmanagement, Protokollierung, Überwachung und Anpassung besprochen, um Plattformadministratoren die Einrichtung und Verwaltung der SageMaker AI Studio Platform zu ermöglichen.

Anhang

Vergleich von Mehrmandantenverhältnissen

Tabelle 2 — Vergleich mehrerer Mandanten

Mehrere Domains	Mehrere Konten	Attributbasierte Zugriffsk ontrolle (ABAC) innerhalb einer einzigen Domain
Die Ressourcenisolieru ng wird mithilfe von Tags erreicht. SageMaker AI Studio kennzeichnet automatisch alle Ressourcen mit der Domäne ARN und dem Benutzerprofil/ Bereich. ARN	Jeder Mandant hat sein eigenes Konto, sodass eine absolute Ressourcenisolierung besteht.	Die Ressourcenisolierung wird mithilfe von Tags erreicht. Benutzer müssen das Tagging der erstellten Ressourcen für ABAC verwalten.
Die Liste APIs kann nicht durch Tags eingeschränkt werden. Die UI-Filterung von Ressourcen erfolgt für gemeinsam genutzte Bereiche. Bei API Listenauf rufen, die über AWS CLI oder Boto3 getätigt werden, SDK werden jedoch Ressourcen in der gesamten Region aufgelist et.	Eine APIs Isolierung von Listen ist ebenfalls möglich, da sich die Mandanten in ihren eigenen Konten befinden.	Die Liste APIs kann nicht durch Tags eingeschränkt werden. APIListenanrufe, die über AWS CLI oder den Boto3 SDK getätigt wurden, listet Ressourcen in der gesamten Region auf.
SageMaker Die Rechen- und Speicherkosten von AI Studio pro Mandant können einfach überwacht werden, indem Domain ARN als Kostenzuw eisungs-Tag verwendet wird.	SageMaker Die Rechen- und Speicherkosten von AI Studio pro Mandant lassen sich mit einem speziellen Konto einfach überwachen.	SageMaker Die Rechenkos ten von AI Studio pro Mandant müssen mithilfe benutzerd efinierter Tags berechnet werden.

Mehrere Domains	Mehrere Konten	Attributbasierte Zugriffsk ontrolle (ABAC) innerhalb einer einzigen Domain
		SageMaker Die Speicherk osten von AI Studio können nicht pro Domain überwacht werden, da sich alle Mandanten das gleiche EFS Volumen teilen.
Servicekontingenten werden auf Kontoebene festgelegt, sodass ein einzelner Mandant immer noch alle Ressourcen verbrauchen kann.	Servicekontingenten können auf Kontoebene für jeden Mandanten festgelegt werden.	Servicekontingenten werden auf Kontoebene festgelegt, sodass ein einzelner Mandant immer noch alle Ressourcen verbrauchen kann.
Die Skalierung auf mehrere Mandanten kann über Infrastru cture as Code (IaC) oder Service Catalog erreicht werden.	Die Skalierung auf mehrere Mandanten beinhaltet Organizations und den Verkauf mehrerer Konten.	Für die Skalierung ist für jeden neuen Mandanten eine mandantenspezifische Rolle erforderlich, und Benutzerp rofile müssen manuell mit Mandantennamen versehen werden.
Die Zusammenarbeit zwischen Benutzern innerhalb eines Mandanten ist über gemeinsam genutzte Bereiche möglich.	Die Zusammenarbeit zwischen Benutzern innerhalb eines Mandanten ist über gemeinsam genutzte Bereiche möglich.	Alle Mieter haben Zugriff auf denselben gemeinsamen Raum für die Zusammena rbeit.

SageMaker Sicherung und Wiederherstellung von AI Studio-Domänen

Im Falle eines versehentlichen EFS Löschens oder wenn eine Domain aufgrund von Netzwerk- oder Authentifizierungsänderungen neu erstellt werden muss, folgen Sie diesen Anweisungen.

Option 1: Erstellen Sie eine Sicherungskopie aus einer bestehenden Datei EFS mit EC2

SageMaker Studio-Domain-Backup

1. Listet Benutzerprofile und Bereiche in SageMaker Studio auf (CLI, SDK).
2. Ordnen Sie Benutzerprofile/Bereiche der Option „Ein“ zuUIDs. EFS
   a. Für jeden Benutzer in der Liste von users/spaces, describe the user profile/space (CLI, SDK).
   b. Ordnen Sie das Benutzerprofil/den Bereich zu. `HomeEfsFileSystemUid`
   c. Ordnen `UserSettings['ExecutionRole']` Sie das Benutzerprofil zu, wenn Benutzer unterschiedliche Ausführungsrollen haben.
   d. Identifizieren Sie die standardmäßige Space-Ausführungsrolle.
3. Erstellen Sie eine neue Domäne und geben Sie die standardmäßige Space-Ausführungsrolle an.
4. Erstellen Sie Benutzerprofile und Bereiche.
   - Erstellen Sie für jeden Benutzer in der Benutzerliste mithilfe der Ausführungsrollenzuordnung ein Benutzerprofil (CLI, SDK).
5. Erstellen Sie eine Zuordnung für das neue EFS undUIDs.
   a. Beschreiben Sie für jeden Benutzer in der Benutzerliste das Benutzerprofil (CLI, SDK).
   b. Ordnen Sie das Benutzerprofil zu`HomeEfsFileSystemUid`.
6. Löschen Sie optional alle Apps, Benutzerprofile und Bereiche und löschen Sie dann die Domäne.

EFS-Sicherung

Gehen Sie wie folgt vorEFS, um eine Sicherungskopie zu erstellen:

1. Starten Sie die EC2 Instanz und fügen Sie die Sicherheitsgruppen für eingehende/ausgehende Nachrichten der alten SageMaker Studio-Domäne an die neue EC2 Instanz an (lassen Sie NFS Datenverkehr über TCP Port 2049 zu). Weitere Informationen finden Sie unter Connect von SageMaker Studio-Notebooks unter a VPC mit externen Ressourcen.
2. Hängen Sie das SageMaker EFS Studio-Volume in die neue EC2 Instanz ein. Weitere Informationen finden Sie unter EFS Dateisysteme einhängen.
3. Kopieren Sie die Dateien in den EBS lokalen Speicher: `>sudo cp -rp /efs /studio-backup:`
   a. Hängen Sie die neuen Domänensicherheitsgruppen an die EC2 Instanz an.

b. Hängen Sie das neue EFS Volume in die EC2 Instance ein.

c. Kopieren Sie Dateien auf das neue EFS Volume.

d. Für jeden Benutzer in der Sammlung des Benutzers:

i. Erstellen Sie das Verzeichnis:`mkdir new_uid`.

ii. Kopieren Sie Dateien aus dem alten UID Verzeichnis in das neue UID Verzeichnis.

iii. Ändern Sie den Besitz für alle Dateien: `chown <new_UID>` für alle Dateien.

Option 2: Erstellen Sie EFS mithilfe von S3 und der Lebenszykluskonfiguration eine Sicherungskopie von vorhandenen Daten

1. Weitere Informationen finden Sie unter Migrieren Sie Ihre Arbeit mit Amazon Linux 2 auf eine SageMaker Amazon-Notebook-Instance.
2. Erstellen Sie einen S3-Bucket für Backups (z. `>studio-backup` B.
3. Listet alle Benutzerprofile mit Ausführungsrollen auf.
4. Legen Sie in der aktuellen SageMaker Studio-Domäne ein LCC Standardskript auf Domänenebene fest.
   - Kopieren Sie in der LCC alles in `/home/sagemaker-user` das Benutzerprofilpräfix in S3 (z. B.`s3://studio-backup/studio-user1`).
5. Starten Sie alle Standard-Jupyter Server-Apps neu (LCCdamit sie ausgeführt werden können).
6. Löschen Sie alle Apps, Benutzerprofile und Domänen.
7. Erstellen Sie eine neue SageMaker Studio-Domäne.
8. Erstellen Sie neue Benutzerprofile aus der Liste der Benutzerprofile und Ausführungsrollen.
9. Richten Sie ein LCC auf Domänenebene ein:
   - Kopieren Sie im LCC alles im Benutzerprofilpräfix in S3 nach `/home/sagemaker-user`
10. Erstellen Sie Standard-Jupyter Server-Apps für alle Benutzer mit der LCCKonfiguration (CLI,). SDK

SageMaker Studio-Zugriff mithilfe von Assertion SAML

Einrichtung der Lösung:

1. Erstellen Sie eine SAML Anwendung in Ihrem externen IdP.

2. Richten Sie den externen IdP als Identitätsanbieter in IAM ein.
3. Erstellen Sie eine `SAMLValidator` Lambda-Funktion, auf die der IdP zugreifen kann (über eine Funktion URL oder ein API Gateway).
4. Erstellen Sie eine `GeneratePresignedUrl` Lambda-Funktion und ein API Gateway, um auf die Funktion zuzugreifen.
5. Erstellen Sie eine IAM Rolle, die Benutzer übernehmen können, um das API Gateway aufzurufen. Diese Rolle sollte als SAML Assertion als Attribut im folgenden Format übergeben werden:
   - Attributname: https://aws.amazon.com/SAML/ Attribute/Rolle
   - Attributwert:, `<IdentityProviderARN> <RoleARN>`
6. Aktualisieren Sie den Endpunkt SAML Assertion Consumer Service (ACS) auf den `SAMLValidator` URL Invoke.

SAMLValidator-Beispielcode:

```
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
```

Weitere Informationen

- Einrichtung sicherer, gut verwalteter Umgebungen für maschinelles Lernen auf AWS (AWS Blog)
- Konfiguration von Amazon SageMaker AI Studio für Teams und Gruppen mit vollständiger Ressourcenisolierung (AWS Blog)
- Einführung in Amazon SageMaker AI Studio mit AWS SSO Okta Universal Directory (Blog)AWS
- So konfigurieren Sie SAML 2.0 für AWS Account Federation (Okta-Dokumentation)
- Erstellen Sie eine sichere Plattform für Machine Learning für Unternehmen auf AWS (AWS technischer Leitfaden)
- Passen Sie Amazon SageMaker AI Studio mithilfe von Lifecycle-Konfigurationen an (AWS Blog)
- Bringen Sie Ihr eigenes benutzerdefiniertes Container-Image in Amazon SageMaker AI Studio-Notizbücher (AWS Blog)
- Erstellen Sie benutzerdefinierte SageMaker KI-Projektvorlagen — Best Practices (AWS Blog)
- Implementierung eines Modells mit mehreren Konten mit Amazon SageMaker AI Pipelines (Blog)AWS
- Teil 1: Wie NatWest Group eine skalierbare, sichere und nachhaltige MLOps Plattform aufgebaut hat (Blog)AWS
- Secure Amazon SageMaker AI Studio vorkonfiguriert URLs Teil 1: Grundlegende Infrastruktur (Blog)AWS

Mitwirkende

Zu den Mitwirkenden an diesem Dokument gehören:

- Ram Vittal, Architekt für ML-Lösungen, Amazon Web Services
- Sean Morgan, Architekt für ML-Lösungen, Amazon Web Services
- Durga Sury, Architektin für ML-Lösungen, Amazon Web Services

Besonderer Dank geht an die folgenden Personen, die Ideen, Überarbeitungen und Perspektiven beigesteuert haben:

- Alessandro Cerè, Architekt für KI/ML-Lösungen, Amazon Web Services
- Sumit Thakur, SageMaker KI-Produktleiter, Amazon Web Services
- Han Zhang, leitender Softwareentwicklungsingenieur, Amazon Web Services
- Bhadrinath Pani, Softwareentwicklungsingenieur, Amazon Web Services, Amazon Web Services

Dokumentversionen

Abonnieren Sie den RSS-Feed, um über Aktualisierungen dieses Whitepapers informiert zu werden.

Änderung	Beschreibung	Datum
Das Whitepaper wurde aktualisiert	Defekte Links wurden behoben und zahlreiche redaktionelle Änderungen wurden durchgehend vorgenommen.	25. April 2023
Erstveröffentlichung	Whitepaper veröffentlicht.	19. Oktober 2022

Hinweise

Die Kunden sind dafür verantwortlich, die Informationen in diesem Dokument selbst unabhängig zu beurteilen. Dieses Dokument: (a) dient nur zu Informationszwecken, (b) stellt aktuelle AWS Produktangebote und Praktiken dar, die ohne vorherige Ankündigung geändert werden können, und (c) stellt keine Verpflichtungen oder Zusicherungen von AWS und seinen verbundenen Unternehmen, Lieferanten oder Lizenzgebern dar. AWSProdukte oder Dienstleistungen werden „wie sie sind“ ohne ausdrückliche oder stillschweigende Garantien, Zusicherungen oder Bedingungen jeglicher Art bereitgestellt. Die Verantwortlichkeiten und Verbindlichkeiten AWS gegenüber seinen Kunden werden durch AWS Vereinbarungen geregelt, und dieses Dokument ist weder Teil einer Vereinbarung zwischen AWS und seinen Kunden noch ändert es diese.

AWS-Glossar

Die neueste AWS-Terminologie finden Sie im AWS-Glossar in der AWS-Glossar-Referenz.

Die vorliegende Übersetzung wurde maschinell erstellt. Im Falle eines Konflikts oder eines Widerspruchs zwischen dieser übersetzten Fassung und der englischen Fassung (einschließlich infolge von Verzögerungen bei der Übersetzung) ist die englische Fassung maßgeblich.