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 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 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 bewährten Methoden sind für die Bereitstellung von SageMaker 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 AWSWell-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 Studio als Ihre ML-Plattform verwalten, benötigen Sie Anleitungen zu bewährten Methoden, damit Sie fundierte Entscheidungen treffen können, damit Sie Ihre ML-Plattform bei wachsenden Workloads skalieren können. 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 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 gewährleisten.
- 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, Instanztypen, Apps und 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 Studio-Notebook-Erlebnis mit Ihren eigenen Images 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 Produktionsanwendungen 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

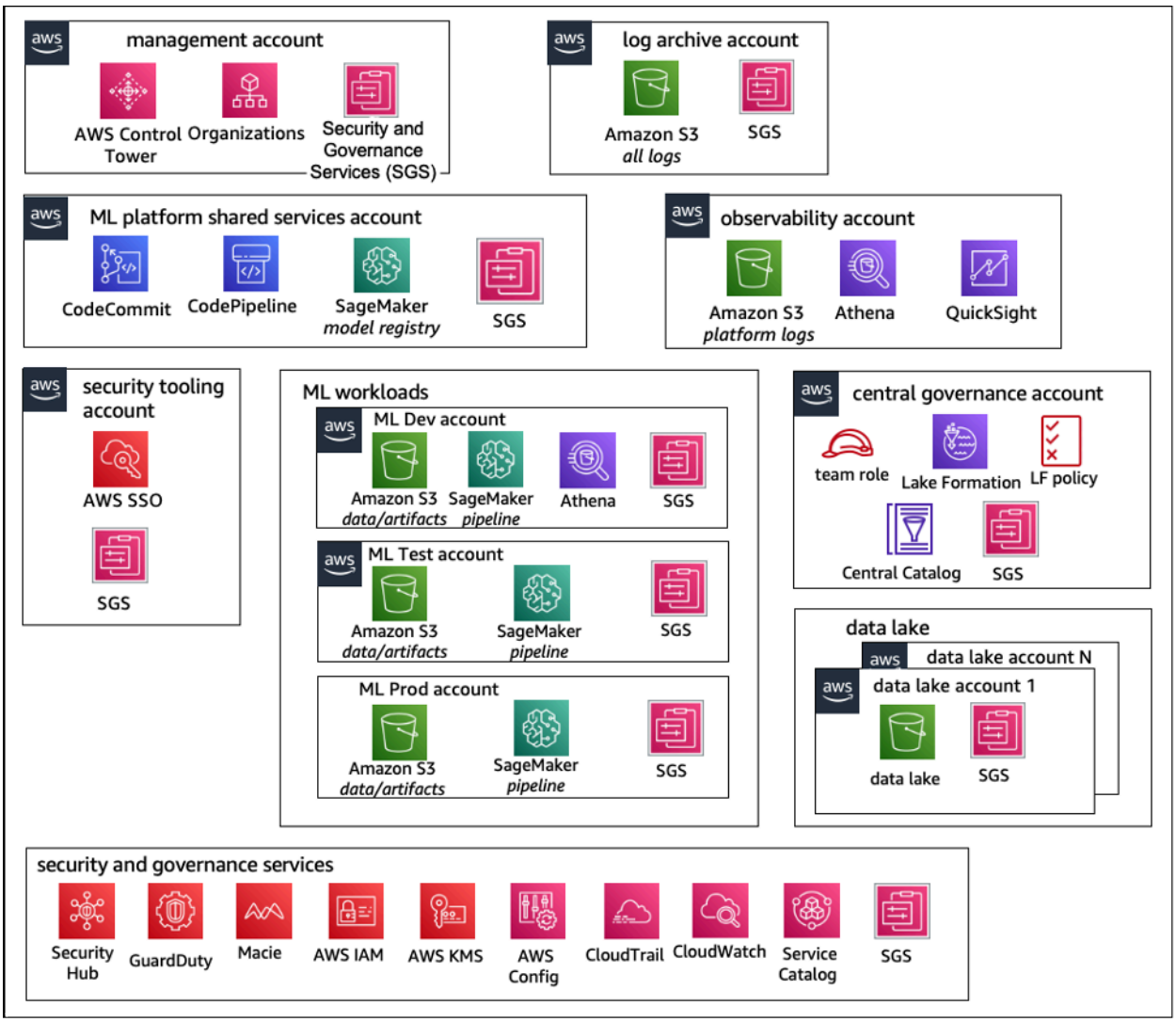
In diesem Abschnitt stellen wir kurz eine Kontostruktur nach dem Betriebsmodell vor, mit der Sie beginnen und die Sie entsprechend den betrieblichen Anforderungen Ihres Unternehmens ändern können. Unabhängig davon, für welches Betriebsmodell Sie sich entscheiden, empfehlen wir die Implementierung der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

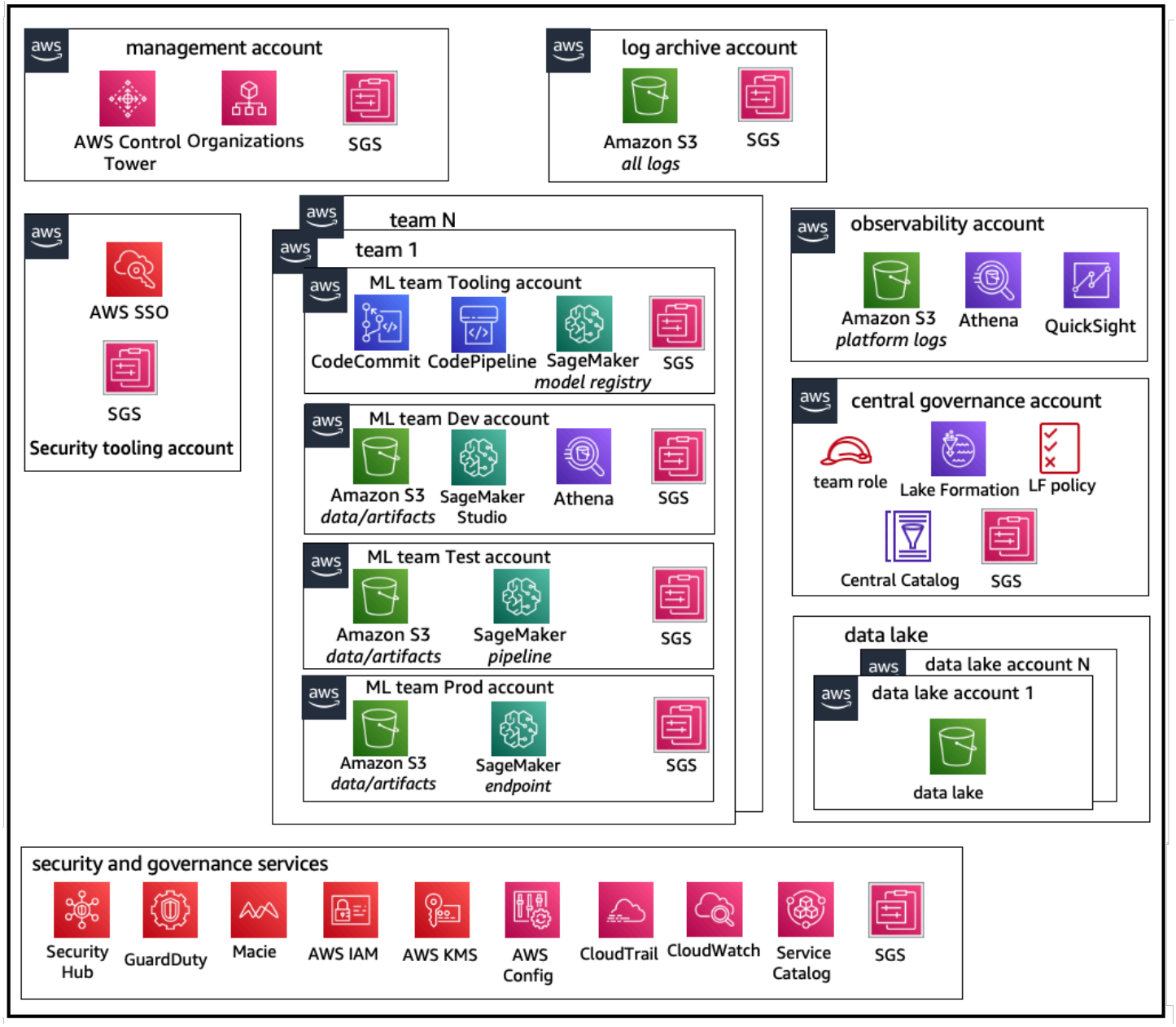
- Ein Shared-Services-Tooling-Konto, das die Anforderungen von Machine Learning Operations (MLOps) in allen Data-Science-Teams erfüllt.
- Konten für die Entwicklung, den Test und die Produktion von ML-Workloads, die von allen Data-Science-Teams gemeinsam genutzt werden.
- Governance-Richtlinien, um sicherzustellen,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



Kontostruktur eines zentralisierten Betriebsmodells

Dezentrale Modellkontenstruktur

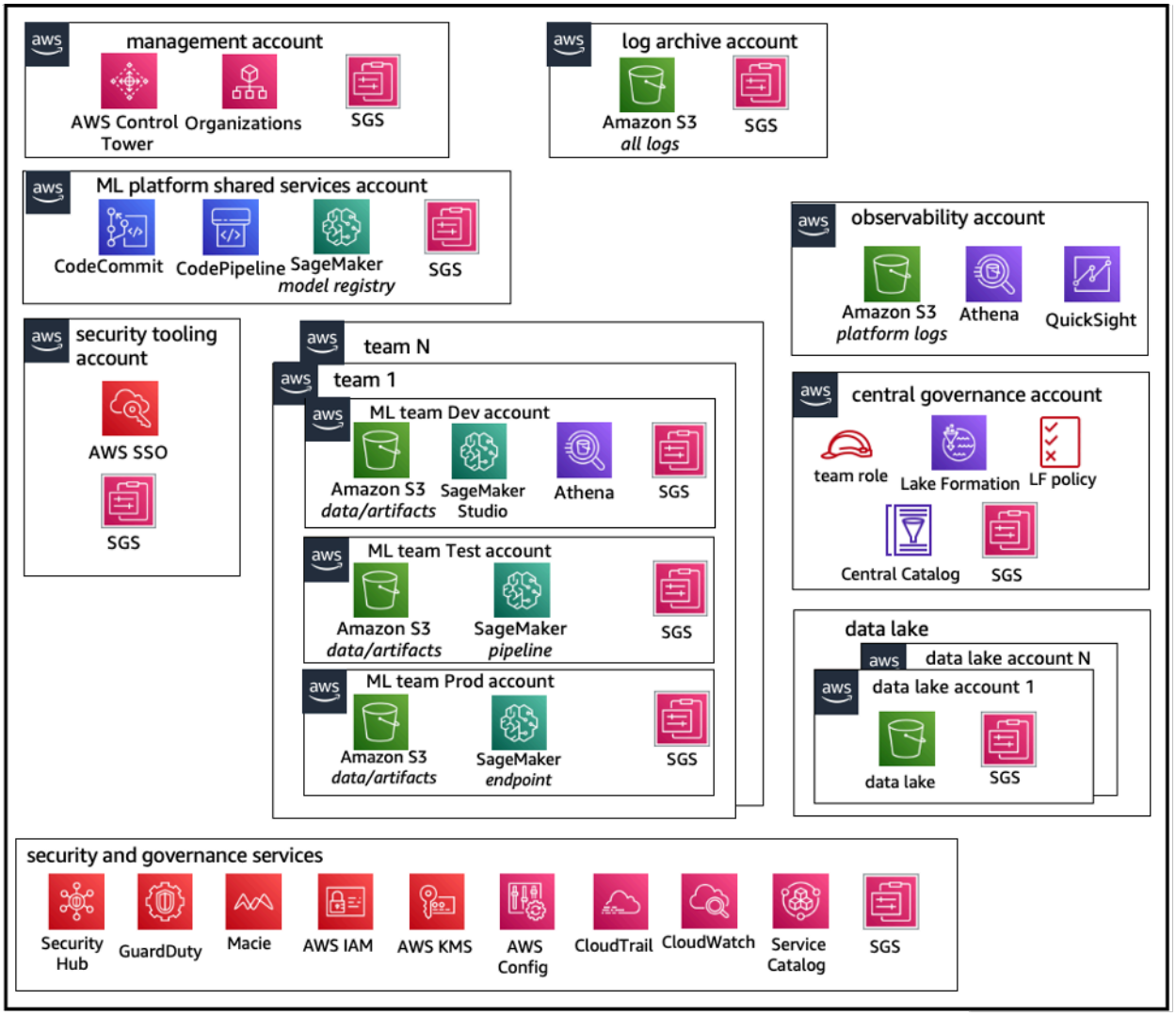
In diesem Modell arbeitet jedes ML-Team unabhängig bei der Bereitstellung, Verwaltung und Verwaltung von ML-Konten und -Ressourcen. Wir empfehlen ML-Teams jedoch, einen zentralisierten Ansatz für Beobachtbarkeit und Datenverwaltung zu verwenden, um die Datenverwaltung und das Auditmanagement zu vereinfachen.



Kontostruktur eines dezentralen Betriebsmodells

Kontostruktur nach föderiertem Modell

Dieses Modell ähnelt dem zentralisierten Modell. Der Hauptunterschied besteht jedoch darin, dass jedes Data-Science-/ML-Team seine eigenen Workload-Konten für Entwicklung/Test/Produktion erhält, die eine robuste physische Isolierung seiner ML-Ressourcen ermöglichen und es jedem Team ermöglichen, 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 Studio-Instanz (z. B. SageMakerDomain) zu implementieren, aber wägen Sie diese Vorteile gegen Kompromisse wie Explosionsradius, Kostenzuweisung und Beschränkungen auf Kontoebene ab, wenn Sie mehrere

Teams in einer einzigen Studio-Domain zusammenführen. SageMaker In den folgenden Abschnitten erfahren Sie mehr über diese Kompromisse und bewährte Methoden.

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

SageMaker Ressourcen, die aus einer Domain erstellt wurden, werden automatisch mit dem Amazon Resource Name (ARN) der Domain und dem Benutzerprofil oder Space-ARN versehen, 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

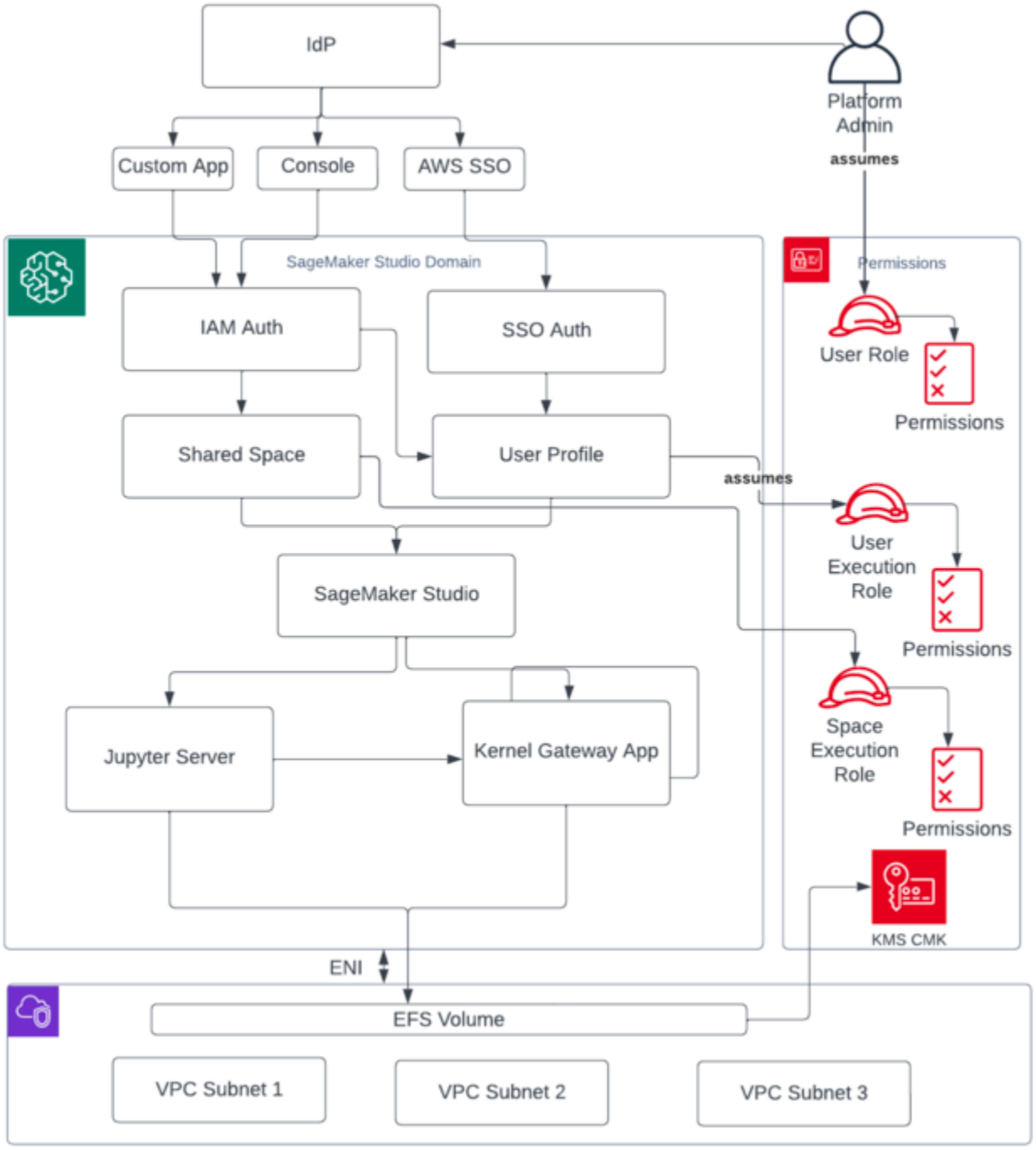
Schließlich können Sie mithilfe von. eine Self-Service-Bereitstellung von SageMaker Studio-Ressourcen für mehrere Konten implementieren. AWS Service Catalog Weitere Informationen finden Sie unter AWS Service CatalogProdukte in mehreren AWS-Konten und AWS-Regionen verwalten.

Domänenverwaltung

Eine SageMaker Amazon-Domain besteht aus:

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

Das folgende Diagramm bietet einen allgemeinen Überblick über die verschiedenen Komponenten, aus denen eine SageMakerStudio Domain besteht:

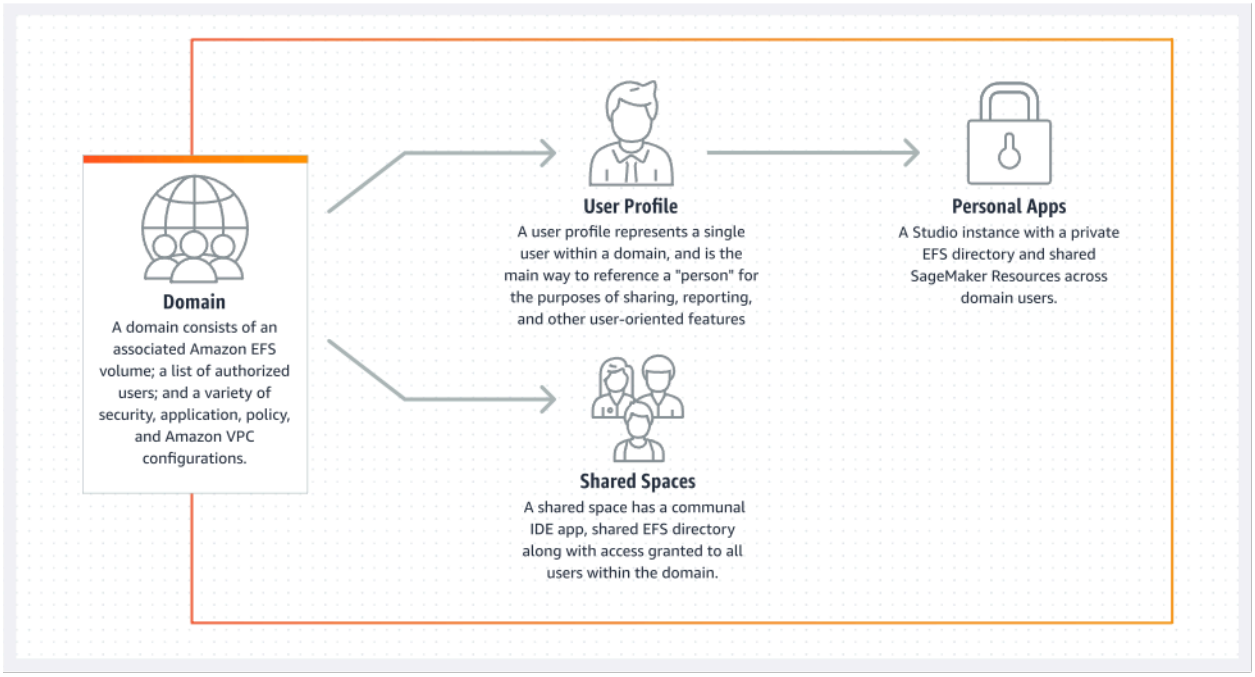


Überblick über verschiedene Komponenten, aus denen eine SageMaker Studio-Domäne besteht

Mehrere Domains und gemeinsam genutzte Bereiche

Amazon unterstützt SageMaker jetzt die Erstellung mehrerer SageMaker Domains in einer einzigen AWS-Region für jedes Konto. Jede Domäne kann ihre eigenen Domäneneinstellungen wie den Authentifizierungsmodus und Netzwerkeinstellungen wie VPC und 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 Domain, die im IAM-Authentifizierungsmodus eingerichtet ist, kann gemeinsam genutzten Speicherplatz für die Zusammenarbeit zwischen Benutzern nahezu in Echtzeit nutzen. Mit einem gemeinsamen Bereich erhalten Benutzer Zugriff auf ein gemeinsam verwendetes Amazon EFS-Verzeichnis und eine gemeinsam genutzte JupyterServerApp für die Benutzeroberfläche und können diese 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 dieselbe IDE benötigen. Der Benutzer kann nahezu in Echtzeit auf seine Notizbücher zugreifen, sie lesen, bearbeiten und teilen, was ihm den schnellsten Weg gibt, 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 einen AWS Identity and Access Management (IAM-) Verbund für Ihre SageMaker 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 Studio mit der IAM-Option finden Sie unter Onboard to Amazon SageMaker Domain Using IAM Identity Center.

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

Um den SSO-Verbund (Single Sign-On) zu verwenden, müssen Sie ihn AWS IAM Identity Center in Ihrem AWS OrganizationsVerwaltungskonto in derselben Region aktivieren, in der Sie Studio ausführen SageMaker 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 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 Inhalte zu teilen, Berichte zu erstellen und andere benutzerorientierte Funktionen zu nutzen. Diese Entität wird erstellt, wenn ein Benutzer Studio einloggt. toSageMaker 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 (Amazon EFS) -Stammverzeichnis des Benutzers. Wir empfehlen, für jeden physischen Benutzer der SageMaker Studio-Anwendung ein Benutzerprofil zu erstellen. Jeder Benutzer hat sein eigenes Verzeichnis auf Amazon EFS, und Benutzerprofile können nicht domänenübergreifend in demselben Konto gemeinsam genutzt werden.

Jedes Benutzerprofil, das sich die SageMaker Studio-Domain teilt, erhält dedizierte Rechenressource (n) (wie SageMaker Amazon Elastic Compute Cloud (Amazon EC2) -Instance (s)), um Notebooks auszuführen. 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 Studio-Notebook für einen Benutzer starten, indem Sie auf die vorsignierte URL zugreifen oder sich mit AWS IAM IDC anmelden, wird die Jupyter Server-App in der vom Service verwalteten VPC-Instance gestartet. SageMaker Jeder Benutzer erhält seine eigene dedizierte Jupyter Server-App in einer privaten App. Standardmäßig wird die Jupyter Server-App für SageMaker 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 Studio-Schnittstelle erstellt werden und läuft auf dem ausgewählten Instanztyp. Diese App kann mit einem der integrierten SageMaker Studio-Images ausgeführt werden, die mit gängigen Data Science- und Deep-Learning-Paketen wie TensorFlowApache MXNet und vorkonfiguriert sind. PyTorch

Benutzer können mehrere Jupyter-Notebook-Kernel, Terminalsitzungen und interaktive Konsolen innerhalb derselben Studio Image/Kernel Gateway-App starten und ausführen. SageMaker Benutzer können auch bis zu vier Kernel-Gateway-Apps oder -Images auf derselben physischen Instanz ausführen — jede davon isoliert durch ihren 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 auf derselben Instanz sowohl ein einfaches Notebook mit dem integrierten Data-Science-Image von SageMaker Studio als auch ein anderes Notebook mit dem integrierten TensorFlow Image ausführen. Benutzern wird die Zeit in Rechnung gestellt, in der die Instanz ausgeführt wird. Um Kosten zu vermeiden, wenn der Benutzer SageMaker 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 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-Skripts verwenden, um die eigenen Pakete des Benutzers in SageMaker Studio zu übertragen und sie über Instanzwechsel und das Starten neuer Instanzen beizubehalten.

Amazon Elastic File System-Volume

Wenn eine Domain erstellt wird, wird ein einzelnes Amazon Elastic File System (Amazon EFS) -Volume erstellt, das von allen Benutzern innerhalb der Domain verwendet werden kann. Jedes Benutzerprofil erhält ein privates Home-Verzeichnis innerhalb des Amazon EFS-Volumes zum Speichern der Notizbücher, GitHub Repositorys und Datendateien des Benutzers. Jeder Bereich innerhalb einer Domain erhält ein privates Verzeichnis innerhalb des Amazon EFS-Volumes, auf das mehrere Benutzerprofile zugreifen können. Der Zugriff auf die Ordner ist durch Dateisystemberechtigungen nach Benutzern getrennt. SageMaker Studio erstellt eine globale eindeutige Benutzer-ID für jedes Benutzerprofil oder jeden Bereich und wendet sie als POSIX-Benutzer-/Gruppen-ID (Portable Operating System Interface) für das Home-Verzeichnis des Benutzers auf EFS an, wodurch verhindert wird, dass andere Benutzer/Bereiche auf seine Daten zugreifen.

Sicherung und Wiederherstellung

Ein vorhandenes EFS-Volume kann nicht an eine neue SageMaker Domäne angehängt werden. Stellen Sie in einer Produktionsumgebung sicher, dass das Amazon EFS-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 Studio-Domäne entfernen und neu erstellen. Der Prozess läuft folgendermaßen ab:

Sichern Sie die Liste der Benutzerprofile, Bereiche und der zugehörigen EFS-Benutzer-IDs (UIDs) über die `DescribeSpace` API-Aufrufe `ListUserProfilesDescribeUserProfile`,`List Spaces`,, und.

1. Erstellen Sie eine neue SageMaker 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 Studio-Domain.

Detaillierte Anweisungen finden Sie im Anhang, Abschnitt Sicherung und Wiederherstellung von SageMaker Studio-Domänen.

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

Amazon EBS-Volume

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

Sicherung des Zugriffs auf die vorsignierte URL

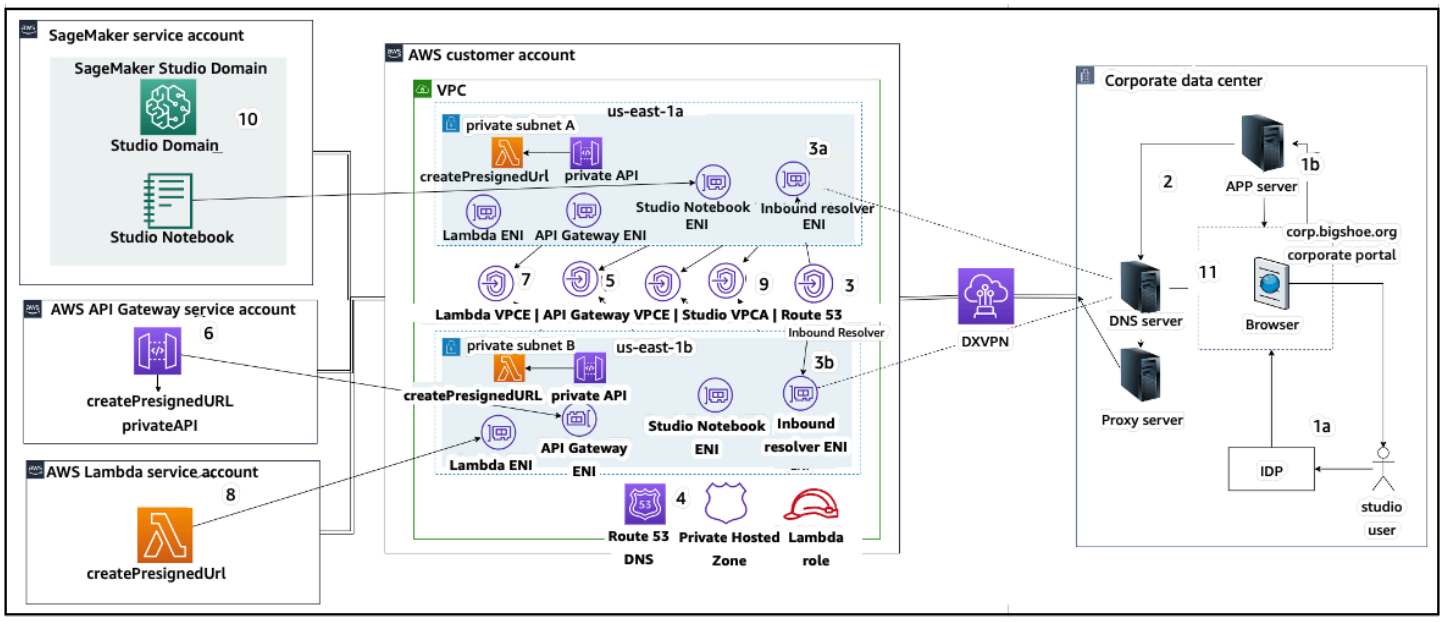
Wenn ein SageMaker Studio-Benutzer den Notizbuch-Link öffnet, validiert SageMaker Studio die IAM-Richtlinie des Verbundbenutzers, um den Zugriff zu autorisieren, und generiert und löst die vorsignierte URL für den Benutzer auf. Da die SageMaker Konsole auf einer Internetdomäne ausgeführt wird, ist diese generierte, vorsignierte URL 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 Datendiebstahl mit vorsignierten URLs:

- Client-IP-Validierung mithilfe der IAM-Richtlinienbedingung `aws:sourceIp`
- Client-VPC-Validierung mithilfe der IAM-Bedingung `aws:sourceVpc`
- Validierung von Client-VPC-Endpunkten mithilfe der IAM-Richtlinienbedingung `aws:sourceVpce`

Wenn Sie von der SageMaker Konsole aus auf SageMaker 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 für das 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 Client-VPC-Endpunktvalidierung mithilfe der IAM-Richtlinienbedingung zu verwenden`aws:sourceVpce`, muss die Erstellung einer vorsignierten URL von derselben Kunden-VPC ausgehen, in der SageMaker Studio bereitgestellt wird, und die Auflösung der vorsignierten URL muss über einen SageMaker Studio-VPC-Endpunkt auf der Kunden-VPC erfolgen. Diese Auflösung der vorsignierten URL während der Zugriffszeit für Benutzer des Unternehmensnetzwerks kann mithilfe von DNS-Weiterleitungsregeln (sowohl in Zscaler als auch im Unternehmens-DNS) und dann mithilfe eines Amazon Route 53-Inbound-Resolvers in den VPC-Endpunkt des Kunden erfolgen, wie in der folgenden Architektur dargestellt:



Zugreifen auf die vorsignierte Studio-URL mit VPC-Endpunkt über das Unternehmensnetzwerk

step-by-step Anleitungen zur Einrichtung der vorherigen Architektur finden Sie unter Sichere vorsignierte URLs von Amazon SageMaker Studio, Teil 1: Grundlegende Infrastruktur.

SageMaker Domain-Kontingente und Limits

- SageMaker Der SSO-Verbund für Studio-Domänen 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.
- VPC- und Subnetzkonfiguration können nach dem Erstellen der Domain nicht geändert werden. Sie können jedoch eine neue Domain mit einer anderen VPC- und Subnetzkonfiguration erstellen.
- Der Domänenzugriff kann nach dem Erstellen der Domain nicht zwischen den Modi IAM und SSO umgeschaltet 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 Studio zusammenschließen AWS-Konten und darauf zugreifen SageMaker . 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.

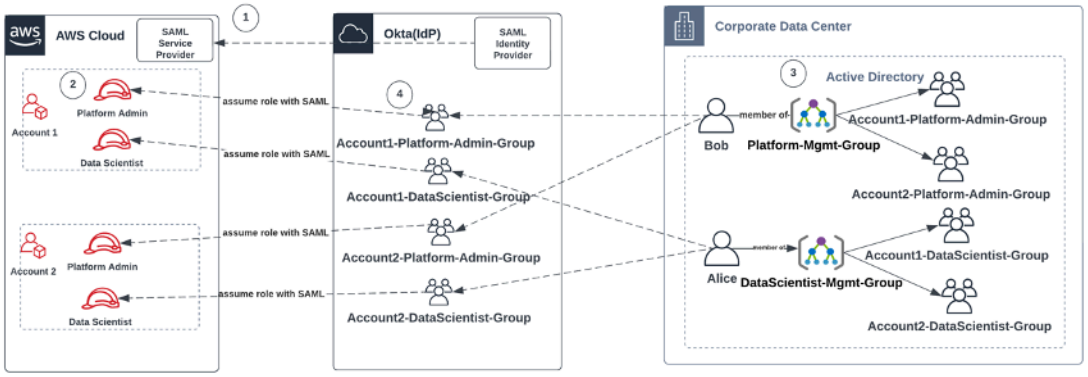
Wie im Abschnitt AWS Security Pillar Identity Management beschrieben, ist es eine bewährte Methode, Ihre Benutzeridentitäten in einem zentralen IdP zu verwalten, da dies die einfache Integration in Ihre Backend-HR-Prozesse erleichtert und Ihnen hilft, den Zugriff auf Ihre Belegschaftsbenutzer zu verwalten.

IdPs wie Okta ermöglicht es Endbenutzern, sich mithilfe von SSO mit Security Assertation Markup Language (SAML) bei einer oder mehreren zu authentifizieren AWS-Konten und Zugriff auf bestimmte Rollen zu erhalten. IdP-Administratoren haben die Möglichkeit, Rollen aus dem AWS-Konten IdP herunterzuladen und diese Benutzern zuzuweisen. Bei der Anmeldung wird Endbenutzern ein AWS Bildschirm angezeigt AWS, auf dem eine Liste der Rollen angezeigt wird, die ihnen in einer oder mehreren AWS Rollen zugewiesen wurden. 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.

In IdP muss für jede spezifische Kombination aus Konto und Rolle, auf die Sie Zugriff gewähren möchten, 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. In AWS sind diese AD-Gruppen IAM-Rollen zugeordnet. Die wichtigsten Schritte des Workflows sind wie folgt:



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

1. AWS Richten Sie unter 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 (z. B. `Account1-Platform-Admin-Group` (auch bekannt als 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 (Konsole) verbunden. AWS SageMaker

Ausführliche Anweisungen zur Einrichtung eines SAML-Kontoverbunds finden Sie in Oktas How to Configure SAML 2.0 for Account Federation. AWS

Benutzerverbund

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

Wenn SageMaker Studio im IAM-Modus eingerichtet ist, greifen SageMaker Studio-Benutzer über eine vorsignierte URL auf die App zu, die einen Benutzer automatisch bei der SageMaker Studio-App anmeldet, wenn der Zugriff über einen Browser erfolgt.

IAM-Benutzer

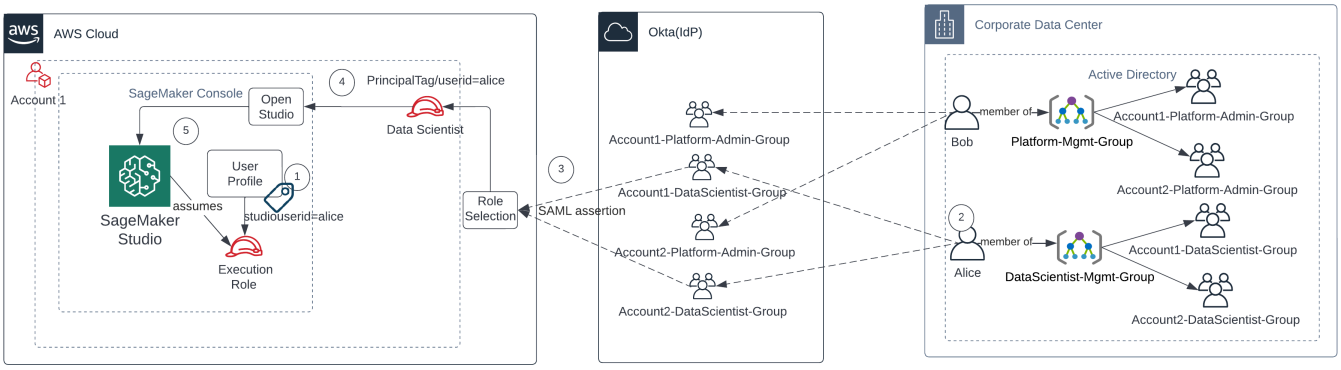
Für IAM-Benutzer erstellt der Administrator SageMaker Studio-Benutzerprofile für jeden Benutzer und ordnet das Benutzerprofil einer IAM-Rolle zu, die die erforderlichen Aktionen ermöglicht, die der Benutzer in Studio ausführen muss. Um zu verhindern, dass ein AWS Benutzer nur auf sein SageMaker Studio-Benutzerprofil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

Die AWS-Konto Verbundmethode ermöglicht es Kunden, sich von ihrem SAML-IdP wie Okta aus mit der SageMaker Konsole zu verbinden. Um zu verhindern, dass Benutzer nur auf ihr Benutzerprofil zugreifen, sollte der Administrator das SageMaker 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 Studio-Benutzerprofil zuzugreifen.



Zugreifen auf SageMaker Studio im IAM-Verbundmodus

1. Das Alice SageMaker 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 Console angemeldet und hat die Rolle des Data Scientist 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 Studio-App-Instanz validiert. Wenn das Profil-Tag gültig ist, wird die SageMaker Studio-App-Instanz gestartet, wobei sie die Ausführungsrolle übernimmt.

Wenn Sie die Erstellung von SageMaker Ausführungsrollen und -richtlinien im Rahmen des Benutzer-Onboardings automatisieren möchten, können Sie dies wie folgt erreichen:

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

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

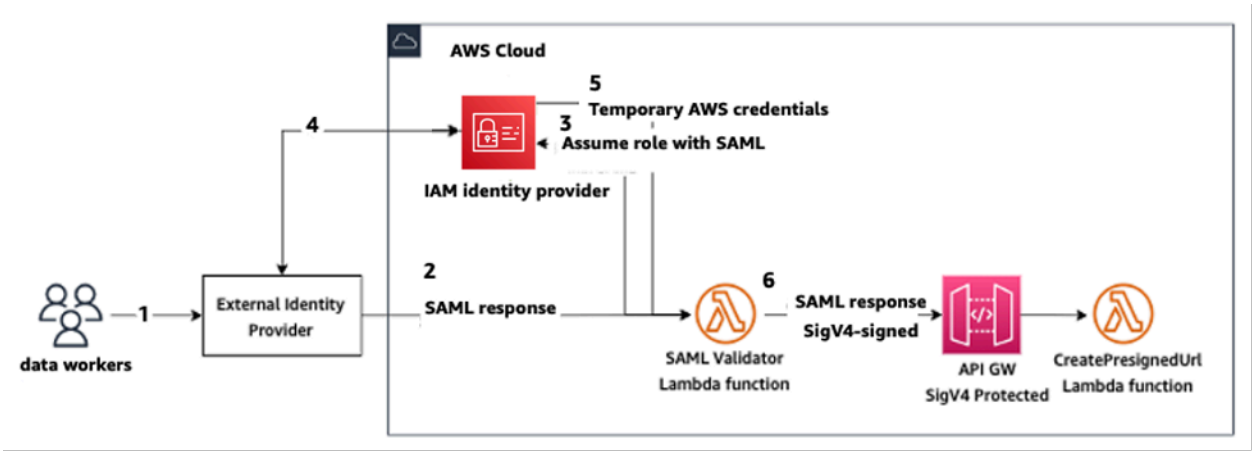
SAML-Authentifizierung mit AWS Lambda

Im IAM-Modus können Benutzer auch mithilfe von SAML-Assertionen in SageMaker 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 zu IAM hinzugefügt. 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-URL zu erstellen.

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

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

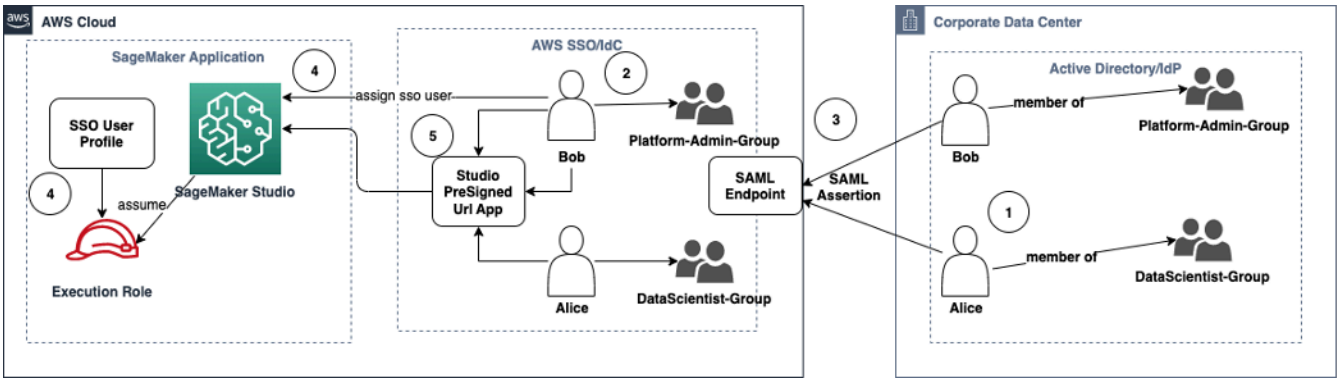
Zusammenfassend lässt sich sagen, dass diese Lösung SageMaker Studio als SAML2.0-Anwendung mit benutzerdefinierter Logik für Authentifizierung und Autorisierung verfügbar macht. Einzelheiten zur Implementierung finden Sie im Anhang, Abschnitt SageMaker Studio-Zugriff mithilfe von SAML-Assertion.



Zugreifen auf SageMaker Studio mit einer benutzerdefinierten SAML-Anwendung

AWS IAM IdC-Verbund

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



Zugreifen auf SageMaker Studio im IAM iDC-Modus

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 dem 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 IdC-SAML-Endpunkt.
4. Im SageMaker Studio ist der IdC-Benutzer der Studio-Anwendung zugewiesen. SageMaker Diese Zuweisung kann mithilfe von IdC Group vorgenommen werden, und SageMaker Studio gilt für

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

5. Der Benutzer greift über die sichere, vorsignierte URL, die als Cloud-Anwendung vom iDC aus gehostet wird, auf die SageMaker Studio-Anwendung zu. SageMaker 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 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 Studio-Benutzeroberfläche zugreifen AWS Management Console und sie aufrufen, können Sie dies tun, indem Sie eine Lambda-Funktion im Backend verwenden, um eine vorsignierte URL für das Benutzerprofil zu generieren und sie zur 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 der UpdateUserProfileAPI aktualisieren.
5. Wenn Sie den Zugriff auf die SageMaker Studio-Benutzeroberfläche im IAM-Modus (mithilfe der generierten vorsignierten URL) auf einen VPC-Endpunkt beschränken möchten, ohne das Internet zu durchqueren, können Sie einen benutzerdefinierten DNS-Resolver verwenden. Weitere Informationen finden Sie im Blogbeitrag Secure Amazon SageMaker Studio Presigned URLs Part 1: Fundamentale Infrastruktur.

Berechtigungsverwaltung

In diesem Abschnitt werden die bewährten Methoden für die Einrichtung häufig verwendeter IAM-Rollen, -Richtlinien und -Guardrails für die Bereitstellung und den Betrieb der Studio-Domäne 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. Wie SageMaker bei einem verwalteten Dienst müssen Sie auch Service Principals berücksichtigen. Dabei handelt es sich um AWS Dienste, die API-Aufrufe im Namen eines Benutzers 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 IAM-Rollen

Diese Rollen werden ausführlich beschrieben, zusammen mit einigen Beispielen für spezifische IAM-Berechtigungen, die sie benötigen.

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

  Da dieser Schulleiter beispielsweise keine Trainingsjobs ausführt, benötigt er keine Berechtigungen, um SageMaker Schulungs- 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 die Verwendung von. SageMaker AWS Management Console

Diese Benutzerrolle benötigt außerdem bestimmte EC2-Berechtigungen, um die Domain in einer privaten VPC zu starten, 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 Prinzipal handelt es sich um den Benutzer, der sich bei SageMaker Studio anmeldet, die Daten untersucht, Verarbeitungs- und Schulungsaufträge und Pipelines erstellt usw. Die primäre Berechtigung, die der Benutzer benötigt, ist die Erlaubnis, SageMaker Studio zu starten. Die übrigen Richtlinien können von der SageMaker Ausführungsdienst-Rolle verwaltet werden.
- SageMaker Ausführungsdienst-Rolle — Da SageMaker es sich um einen verwalteten Dienst handelt, startet er Jobs im Namen eines Benutzers. Diese Rolle ist häufig die umfassendste, was die erlaubten Berechtigungen angeht, da sich viele Kunden dafür entscheiden, eine einzige Ausführungsrolle zu verwenden, um Trainingsjobs, Verarbeitungsjobs oder Modelhosting-Jobs auszuführen. 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 Studio-Domäne eine Rolle zu. Da Kunden jedoch die Flexibilität benötigen, den verschiedenen Benutzerprofilen in der Domäne 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 SageMakerStudio Domänenausführung auch dem Benutzerprofil zuzuordnen.

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

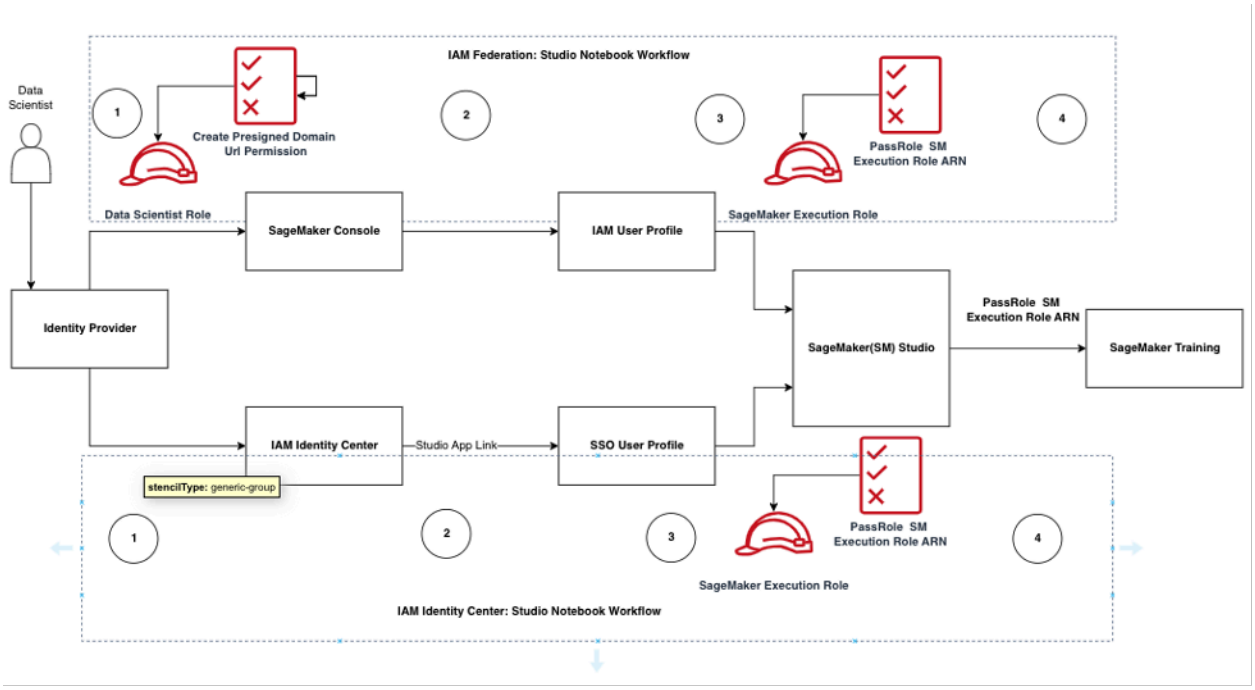
eine individuelle SageMaker Dienstausführungsrolle auf Benutzerebene erstellen. Dabei müssen Sie jedoch Ihre Dienstlimits beachten.

SageMaker Workflow zur Autorisierung von Studio-Notebooks

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

- IAM-Föderation
- IAM Identity Center

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



Der Authentifizierungs- und Autorisierungsablauf für Studio-Benutzer

IAM Federation: SageMaker Studio-Notebook-Workflows

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

Diese Verbundrolle verfügt über `iam:PassRole` API-Berechtigungen für die SageMaker Ausführungsrolle, um die Rolle Amazon Resource Name (ARN) an SageMaker Studio zu übergeben.

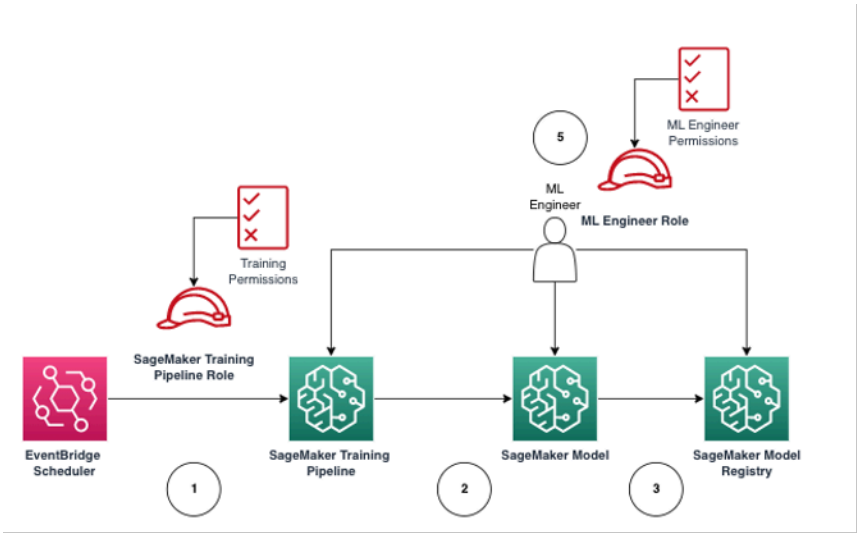
2. Der Data Scientist wählt den Open Studio-Link aus seinem Studio IAM-Benutzerprofil aus, der der SageMaker Ausführungsrolle zugeordnet ist
3. Der SageMaker Studio-IDE-Dienst wird gestartet, wobei die SageMaker Ausführungsrollenberechtigungen des Benutzerprofils vorausgesetzt werden. Diese Rolle verfügt über `iam:PassRole` API-Berechtigungen für die SageMaker Ausführungsrolle, um den Rollen-ARN an den SageMaker Trainingsdienst zu übergeben.
4. Wenn Data Scientist den Trainingsjob in den Remote-Compute-Knoten startet, wird der ARN für die SageMaker Ausführungsrolle an den SageMaker Trainingsdienst übergeben. Dadurch wird eine neue Rollensitzung mit diesem ARN erstellt und der Trainingsjob ausgeführt. Wenn Sie die Berechtigungen für einen Schulungsjob weiter einschränken müssen, können Sie eine trainingsspezifische Rolle erstellen und diese Rolle beim Aufrufen der Schulungs-API den ARN für diese Rolle übergeben.

IAM Identity Center: Arbeitsablauf in SageMaker Studio Notebook

1. Der Data Scientist authentifiziert sich bei seinem Corporate Identity Provider und klickt auf AWS IAM Identity Center. Dem Data Scientist wird das Identity Center-Portal für den Benutzer angezeigt.
2. Der Data Scientist klickt auf den SageMaker Studio-App-Link, der in seinem IdC-Benutzerprofil erstellt wurde und der SageMaker Ausführungsrolle zugeordnet ist.
3. Der SageMaker Studio IDE-Dienst wird gestartet, wobei die SageMaker Ausführungsrollenberechtigungen des Benutzerprofils vorausgesetzt werden. Diese Rolle verfügt über `iam:PassRole` API-Berechtigungen für die SageMaker Ausführungsrolle, um den Rollen-ARN an den SageMaker Trainingsdienst zu übergeben.
4. Wenn der Data Scientist den Trainingsjob in Remote-Compute-Knoten startet, wird der ARN für die SageMaker Ausführungsrolle an den SageMaker Trainingsdienst übergeben. Die Ausführungsrolle ARN erstellt eine neue Rollensitzung mit diesem ARN und führt den Trainingsjob aus. Wenn Sie die Berechtigungen für Schulungsjobs weiter einschränken müssen, können Sie eine trainingsspezifische Rolle erstellen und diese Rolle beim Aufrufen der Schulungs-API den ARN für diese Rolle übergeben.

Bereitgestellte Umgebung: SageMaker 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 Studio Notebooks aus eingeschränkt. In diesem Abschnitt wird erläutert, wie IAM-Rollen mit der SageMaker Trainingspipeline in der bereitgestellten Umgebung zusammenarbeiten.



SageMaker Schulungsablauf in einer verwalteten Produktionsumgebung

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

Datenberechtigungen

Die Fähigkeit von SageMaker Studio-Benutzern, auf jede Datenquelle zuzugreifen, hängt von den Berechtigungen ab, die mit ihrer SageMaker IAM-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 RDS-Datenbanken herzustellen.

Zugriff auf AWS Lake Formation-Daten

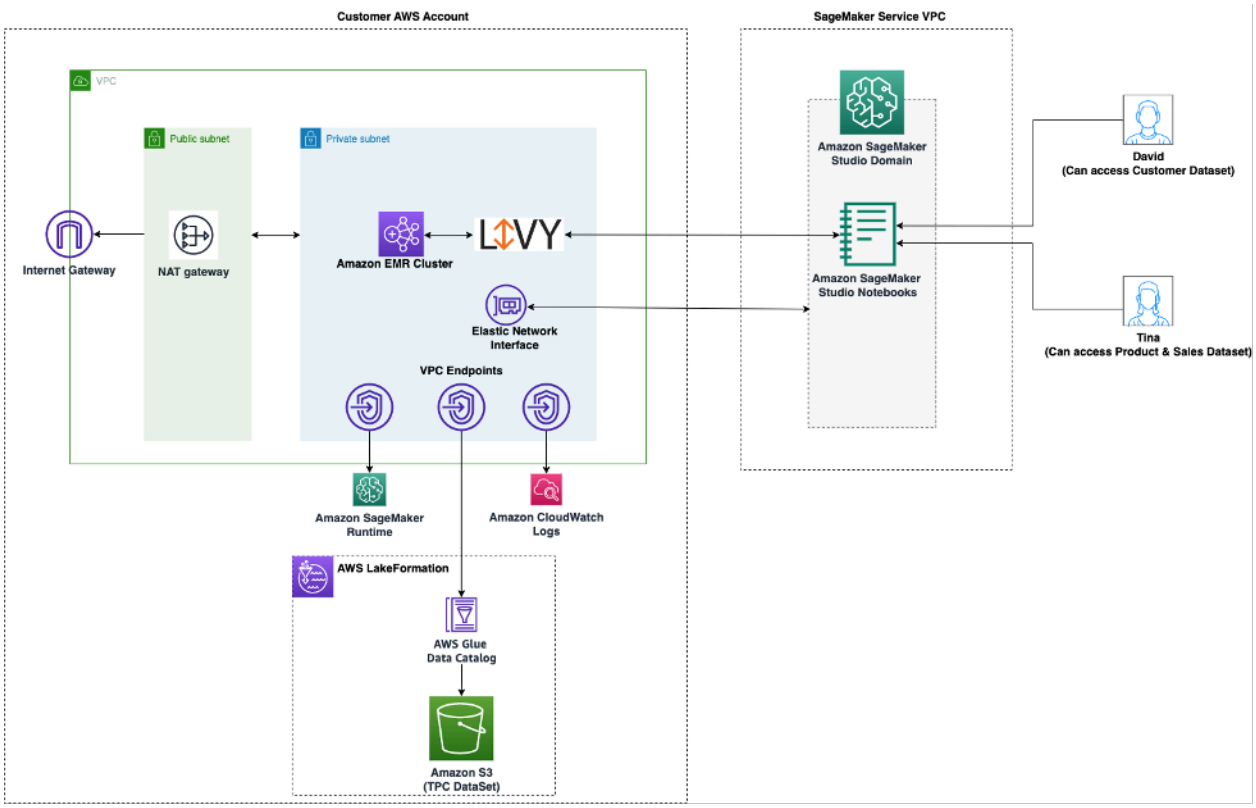
Viele Unternehmen haben damit begonnen, Data Lakes zu nutzen, die von gesteuert werden AWS Lake Formation, um ihren Benutzern einen detaillierten Datenzugriff zu ermöglichen. Als Beispiel für solche verwalteten Daten können Administratoren sensible Spalten für einige Benutzer maskieren und gleichzeitig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

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

1. Von einem SageMaker 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. Das AWS-SDK für 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 SageMaker Studio-Konnektivität zu Amazon EMR, um Daten in großem Umfang zu lesen und zu schreiben. Durch die Verwendung von Apache Livy- und Amazon EMR-Laufzeitrollen verfügt SageMaker Studio über eine native Konnektivität, mit der Sie Ihre SageMaker Ausführungs-IAM-Rolle (oder eine andere autorisierte Rolle) für den Datenzugriff und die Datenverarbeitung an einen Amazon EMR-Cluster übergeben können. up-to-dateAnweisungen finden Sie unter Connect zu einem Amazon EMR-Cluster von Studio aus herstellen.



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

3. Verwenden Sie die native Konnektivität von SageMaker Studio für AWS Glueinteraktive Sitzungen, um Daten in großem Umfang zu lesen und zu schreiben. SageMaker 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 Anwendung von Governance auf Ihre ML-Ressourcen mithilfe von IAM-Richtlinien, Ressourcenrichtlinien, VPC-Endpunktrichtlinien und Service Control Policies (SCPs) beschrieben.

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 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 Studio-Domänen SageMaker

Für SageMaker Studio-Domänen kann die folgende Dienststeuerungsrichtlinie verwendet werden, um den Datenverkehr für den Zugriff auf Kundenressourcen zu erzwingen, sodass dieser nicht über das öffentliche Internet, sondern über die VPC eines Kunden übertragen wird:

```
{
     "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 Bilder SageMaker

Die folgende Richtlinie verhindert, dass ein Benutzer ein nicht autorisiertes SageMaker Bild in seiner Domäne 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 SageMaker VPC-Endpunkte

SageMaker Unterstützt zusätzlich zu VPC-Endpunkten für die SageMaker Steuerungsebene VPC-Endpunkte, mit denen Benutzer eine Verbindung zu SageMakerStudio-Notebooks oder Notebook-Instanzen herstellen können. SageMaker Wenn Sie bereits einen VPC-Endpunkt für eine SageMaker Studio-/Notebook-Instanz eingerichtet haben, erlaubt der folgende IAM-Bedingungsschlüssel nur Verbindungen zu SageMaker Studio-Notebooks, wenn sie über den SageMaker Studio VPC-Endpunkt oder über den API-Endpunkt hergestellt werden. SageMaker

```
{
     "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 Studio-Notebooks auf einen begrenzten IP-Bereich

Unternehmen beschränken den SageMaker Studio-Zugriff 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 Studio-Benutzer auf andere Benutzerprofile zugreifen

Stellen Sie als Administrator beim Erstellen des Benutzerprofils sicher, dass das Profil mit dem SageMaker Studio-Benutzernamen und dem Tag-Schlüssel gekennzeichnet ist`studiouserid`. Der Prinzipal (Benutzer oder Rolle, die dem Benutzer zugewiesen ist) sollte ebenfalls 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 Starten von SageMaker 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 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 Studio-Apps sicher, dass das Benutzerprofil markiert ist. Tags werden automatisch aus dem Benutzerprofil an Apps weitergegeben. Um die Erstellung von Benutzerprofilen mit Tags (unterstützt über CLI und SDK) zu erzwingen, 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. Trainingsjobs und Verarbeitungsjobs, können Sie mithilfe der folgenden Richtlinie Tags verpflichtend 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 Studio

In SageMaker 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-IDs ohne Zugriffsrechte neu zugeordnet. 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-Benutzer erstellen, müssen

Sie sicherstellen, dass die UID/GID des Benutzers mit der UID/GID in der für die benutzerdefinierte App angegebenen identisch ist, die die Konfiguration AppImageConfigfür die Ausführung einer App mit dem benutzerdefinierten Image erstellt. SageMaker 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 dieselbe UID und GID enthalten: `KernelGatewayConfig`

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

Die akzeptablen UID/GID-Werte 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 SageMaker Studio-Notebook-Benutzern keine `DeleteAppImageConfig` Berechtigungen wie `CreateAppImageConfigUpdateAppImageConfig`, oder.

Netzwerkmanagement

Um die SageMaker Studio-Domain einzurichten, müssen Sie das VPC-Netzwerk, Subnetze und Sicherheitsgruppen angeben. Stellen Sie bei der Angabe der VPC und der Subnetze sicher, dass Sie IPs zuweisen, die das Nutzungsvolumen und das erwartete Wachstum berücksichtigen, das in den folgenden Abschnitten beschrieben wird.

VPC-Netzwerkplanung

Kunden-VPC-Subnetze, die der SageMaker Studio-Domain zugeordnet sind, müssen mit dem entsprechenden CIDR-Bereich (Classless Inter-Domain Routing) erstellt werden, abhängig von den folgenden Faktoren:

- Anzahl der Benutzer.
- Anzahl der Apps pro Benutzer.
- Anzahl der eindeutigen Instance-Typen pro Benutzer.
- Durchschnittliche Anzahl von Trainings-Instances pro Benutzer.
- Erwarteter Wachstumsprozentsatz.

SageMaker und teilnehmende AWS Services fügen Elastic Network Interfaces (ENI) für die folgenden Anwendungsfälle in das Kunden-VPC-Subnetz ein:

- Amazon EFS injiziert eine ENI für ein EFS-Mountingziel für die SageMaker Domain (eine IP pro Subnetz/Verfügbarkeitszone, die an die SageMaker Domain angefügt ist).
- SageMaker Studio fügt eine ENI für jede eindeutige Instance ein, die von einem Benutzerprofil oder einem gemeinsam genutzten Bereich verwendet wird. Beispielsweise:
  - Wenn ein Benutzerprofil eine standardmäßige Jupyter-Server-App (eine „System“-Instance), eine Data Science-App und eine Base Python-App (die beide auf einer `ml.t3.medium` Instance ausgeführt werden) ausführt, fügt Studio zwei IP-Adressen ein.
  - Wenn ein Benutzerprofil eine standardmäßige Jupyter-Server-App (eine „System“-Instance), eine Tensorflow-GPU-App (auf einer `ml.g4dn.xlarge` Instance) und eine Data Wrangler-App (auf einer `ml.m5.4xlarge` Instance) ausführt, fügt Studio drei IP-Adressen ein.
- Für jeden VPC-Endpunkt in allen Domain-VPC-Subnetzen/Availability Zones wird eine ENI injiziert (vier IPs für SageMaker VPC-Endpunkte; ~6 IPs für teilnehmende Services-VPC-Endpunkte wie S3, ECR und CloudWatch).

- Wenn SageMaker Schulungs- und Verarbeitungsaufträge mit derselben VPC-Konfiguration gestartet werden, benötigt jeder Auftrag zwei IP-Adressen pro Instance.

> **Note**
>
> VPC-Einstellungen für SageMaker Studio, wie Subnetze und reinen VPC-Datenverkehr, werden nicht automatisch an die von SageMaker Studio erstellten Trainings-/Verarbeitungsaufträge übergeben. Der Benutzer muss beim Aufrufen der Create*Job-APIs nach Bedarf VPC-Einstellungen und Netzwerkisolierung einrichten. Weitere Informationen finden Sie unter Ausführen von Trainings- und Inferenzcontainern im internetfreien Modus.

Szenario: Datenwissenschaftler führt Experimente auf zwei verschiedenen Instance-Typen durch

In diesem Szenario wird davon ausgegangen, dass eine SageMaker Domain im reinen VPC-Datenverkehrsmodus eingerichtet ist. Es sind VPC-Endpunkte wie API, SageMaker Laufzeit, Amazon S3 und Amazon ECR eingerichtet SageMaker.

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

Angenommen, der Datenwissenschaftler führt gleichzeitig einen Schulungsauftrag mit derselben VPC-Konfiguration auf einer `ml.m5.4xlarge` Instance aus.

In diesem Szenario fügt der SageMaker Studio-Service ENIs wie folgt ein:

Tabelle 1 – ENIs, die für ein Experimentierungsszenario in die Kunden-VPC eingefügt werden

Entität	Ziel	ENI injiziert	Hinweise	Level
EFS-Mount ingziel	VPC-Subnetze	Drei	Drei AZs/Subnetze	Domain
VPC-Endpunkte	VPC-Subnetze	30	Drei AZs/Subnetze mit jeweils 10 VPCE	Domain

Entität	Ziel	ENI injiziert	Hinweise	Level
Jupyter Server	VPC-Subnetz	One	Eine IP pro Instance	Benutzer
KernelGateway App	VPC-Subnetz	Zwei	Eine IP pro Instance-Typ	Benutzer
Training	VPC-Subnetz	Zwei	Zwei IPs pro Trainings-Instance Fünf IPs pro Trainings-Instance, wenn EFA verwendet wird	Benutzer

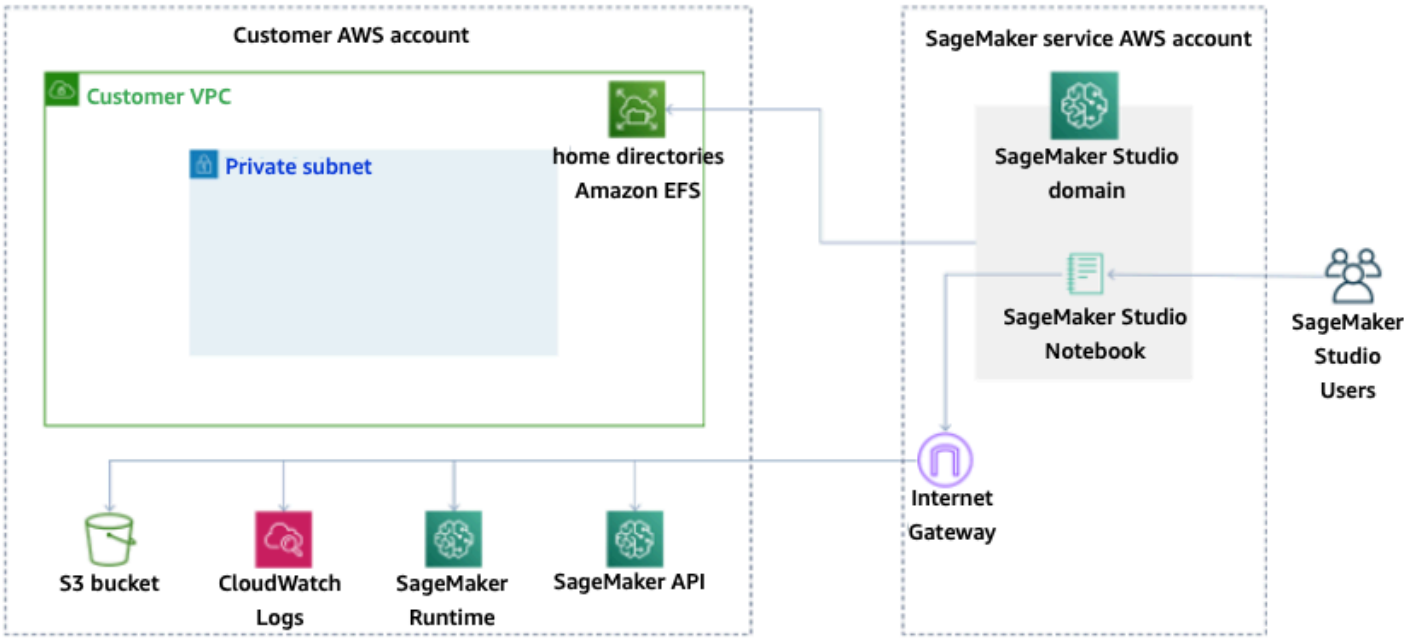
In diesem Szenario werden insgesamt 38 IPs in der Kunden-VPC verbraucht, wobei 33 IPs für alle Benutzer auf Domänenebene und fünf IPs auf Benutzerebene freigegeben werden. Wenn Sie 100 Benutzer mit ähnlichen Benutzerprofilen in dieser Domäne haben, die diese Aktivitäten gleichzeitig ausführen, verbrauchen Sie fünf x 100 = 500 IPs auf Benutzerebene, zusätzlich zum IP-Verbrauch auf Domänenebene, was 11 IPs pro Subnetz entspricht, also insgesamt 511 IPs. In diesem Szenario müssen Sie das VPC-Subnetz-CIDR mit /22 erstellen, das 1024 IP-Adressen zuweist, wobei Platz zum Anwachsen besteht.

VPC-Netzwerkoptionen

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

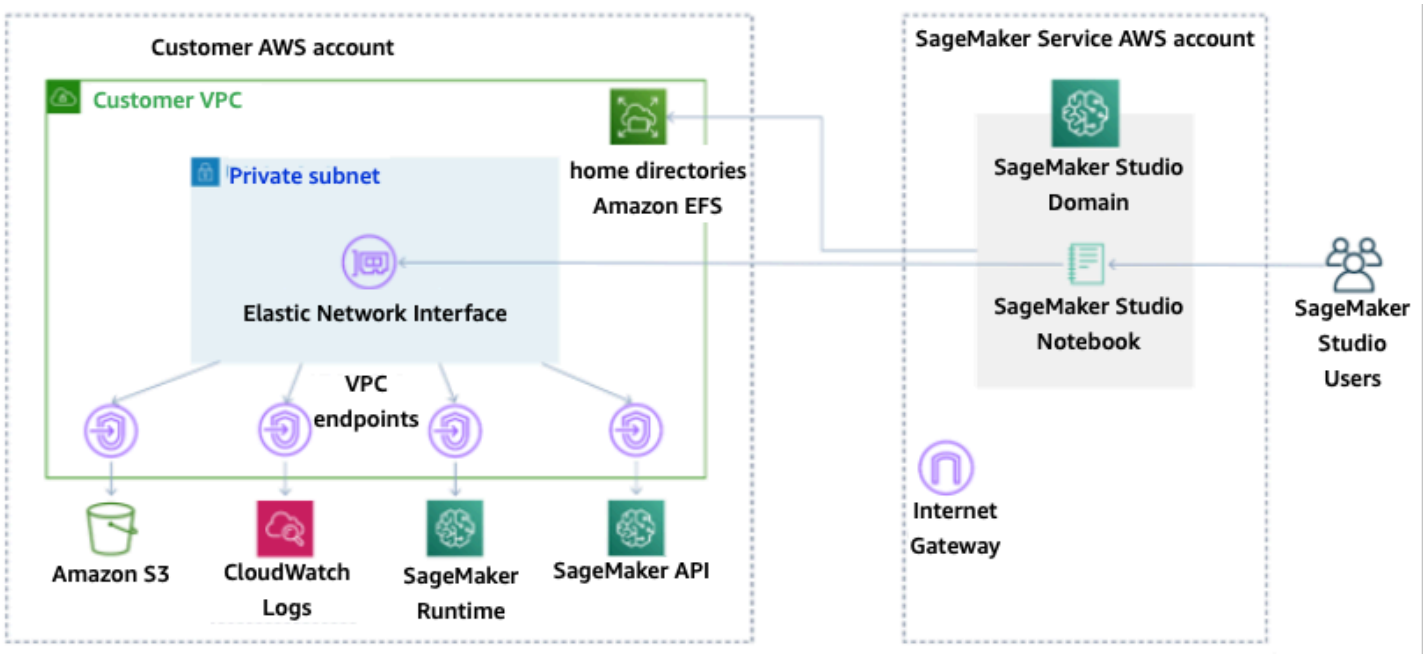
- Nur öffentliches Internet
- Nur VPC

Die Option „Nur öffentliches Internet“ ermöglicht es SageMaker API-Services, das öffentliche Internet über das in der VPC bereitgestellte Internet-Gateway zu nutzen, das SageMaker vom Servicekonto verwaltet wird, wie im folgenden Diagramm dargestellt:



Standardmodus: Internetzugang über SageMaker Servicekonto

Die Option Nur VPC deaktiviert das Internet-Routing von der vom SageMaker Servicekonto verwalteten VPC und ermöglicht es dem Kunden, den Datenverkehr für die Weiterleitung über VPC-Endpunkte zu konfigurieren, wie im folgenden Diagramm dargestellt:



Nur-VPC-Modus: Kein Internetzugang über das SageMaker Servicekonto

Richten Sie für eine im reinen VPC-Modus eingerichtete Domain eine Sicherheitsgruppe pro Benutzerprofil ein, um eine vollständige Isolierung der zugrunde liegenden Instances sicherzustellen. Jede Domain in einem AWS Konto kann ihre eigene VPC-Konfiguration und ihren eigenen Internetmodus haben. Weitere Informationen zum Einrichten der VPC-Netzwerkkonfiguration finden Sie unter Studio SageMaker -Notebooks in einer VPC mit externen Ressourcen verbinden.

Einschränkungen

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

Zum Schutz Ihrer SageMaker Studio-Notizbücher sowie Ihrer Modellbaudaten und Modellartefakte werden die Notizbücher sowie die SageMaker Ergebnisse von Trainings- und Batch-Transformationsaufträgen verschlüsselt. SageMaker verschlüsselt diese standardmäßig mit dem AWSManaged 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 Ressourcen an, damit er für den kontoübergreifenden Zugriff gemeinsam genutzt werden kann.

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

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

Verschlüsselung im Ruhezustand mit AWS KMS

- Sie können Ihren AWS KMSSchlüssel weitergeben, um ein EBS-Volume zu verschlüsseln, das an Notebooks, Schulungen, Tuning, Batch-Transformationsjobs und Endgeräte angeschlossen ist.
- Wenn Sie keinen KMS-Schlüssel angeben, werden sowohl Betriebssystemvolumes (OS) als auch ML-Datenvolumes mit einem vom System verwalteten KMS-Schlüssel SageMaker verschlüsselt.
- 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 Studio stellt sicher, dass ML-Modellartefakte und andere Systemartefakte bei der Übertragung und im Speicher verschlüsselt werden. Anforderungen an die SageMaker -API und die Konsole werden über eine sichere SSL-Verbindung gesendet. Einige Daten innerhalb des Netzwerks sind während der Übertragung (innerhalb der Service-Plattform) unverschlüsselt. Dies umfasst:

- Kommunikation zwischen der Service-Steuerebene und Schulungsauftrags-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 Schulungszeit erhöhen, vor allem wenn Sie mit verteilten Deep Learning-Algorithmen arbeiten.

Standardmäßig SageMaker führt Amazon Trainingsjobs in einer Amazon VPC aus, um die Sicherheit Ihrer Daten zu gewährleisten. Sie können durch das Konfigurieren einer privaten VPC eine weitere Sicherheitsebene zum Schutz Ihrer Schulungscontainer und Daten hinzufügen. Darüber hinaus können Sie Ihre SageMaker Studio-Domain so konfigurieren, dass sie nur im VPC-Modus ausgeführt wird, und VPC-Endpunkte so einrichten, dass sie den Datenverkehr über ein privates Netzwerk weiterleiten, ohne dass ausgehender Datenverkehr über das Internet übertragen wird.

Leitplanken für den Datenschutz

Verschlüsseln Sie SageMaker Hosting-Volumes im Ruhezustand

Verwenden Sie die folgende Richtlinie, um die Verschlüsselung beim Hosten eines SageMaker Endpunkts für Online-Inferenzen 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 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 Studio-Domain-Speichervolume

Erzwingen Sie die Verschlüsselung des Speichervolumes, das an die Studio-Domäne angehängt ist. Diese Richtlinie erfordert, dass ein Benutzer ein CMK zur Verschlüsselung der an Studio-Domänen angehängten Speichervolumes bereitstellt.

```
{
     "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 Studio-Domäne 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 Domäne erstellt wurde, können Sie den angehängten EFS-Volumespeicher nicht mit einem benutzerdefinierten AWS KMS Schlüssel aktualisieren.
- Sie können Trainings-/Verarbeitungsaufträge 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 Studio mithilfe von Amazon überwachen. Amazon sammelt Rohdaten und verarbeitet sie zu lesbaren Metriken CloudWatch, die nahezu in Echtzeit ablaufen. 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. SageMakerVeröffentlicht Metriken standardmäßig in CloudWatch Logs, und diese Protokolle können mit vom Kunden verwalteten Schlüsseln verschlüsselt werden. AWS KMS

Sie können VPC-Endpunkte auch 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 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 der Aufrufer CloudWatch der Job-APIs für SageMaker Training/Processing/Transform Logs in Ihrem Namen senden kann, benötigt er 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 CPU- und Speicherauslastung der Instanz, die Latenz bei Hosting-Aufrufen usw. Sie können Amazon SNS außerdem so konfigurieren, dass Administratoren über Ereignisse informiert werden, wenn bestimmte Schwellenwerte überschritten werden. Der Nutzer der Trainings- und Verarbeitungs-APIs benötigt die folgenden Berechtigungen:

```
{
     "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 Ihren Compliance-Status zu verbessern, überprüfen Sie alle Ihre APIs mitAWS CloudTrail. Standardmäßig werden alle SageMaker APIs mit protokolliert AWS CloudTrail. Für die Aktivierung CloudTrail benötigen Sie keine zusätzlichen IAM-Berechtigungen.

Alle SageMaker Aktionen, mit Ausnahme von `InvokeEndpoint` und`InvokeEndpointAsync`, werden von den Vorgängen protokolliert CloudTrail und sind in den Vorgängen 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:

- Gibt an, ob die Anfrage mit Root- oder IAM-Benutzer-Anmeldeinformationen von AWS ausgeführt wurde.
- Gibt an, ob die Anforderung mit temporären Sicherheitsanmeldeinformationen für eine Rolle oder einen verbundenen Benutzer gesendet wurde.
- Gibt an, ob die Anforderung aus einem anderen AWS-Service gesendet wurde Ein Beispielereignis finden Sie in der CloudTrail Dokumentation Log SageMaker 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 Studio aus. In einem gemeinsam genutzten Bereich beziehen sich alle Aktionen auf den Space-ARN als Quelle, und Sie können keinen Audit durchführen`sourceIdentity`.

Kostenzuweisung

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

Automatisiertes Tagging

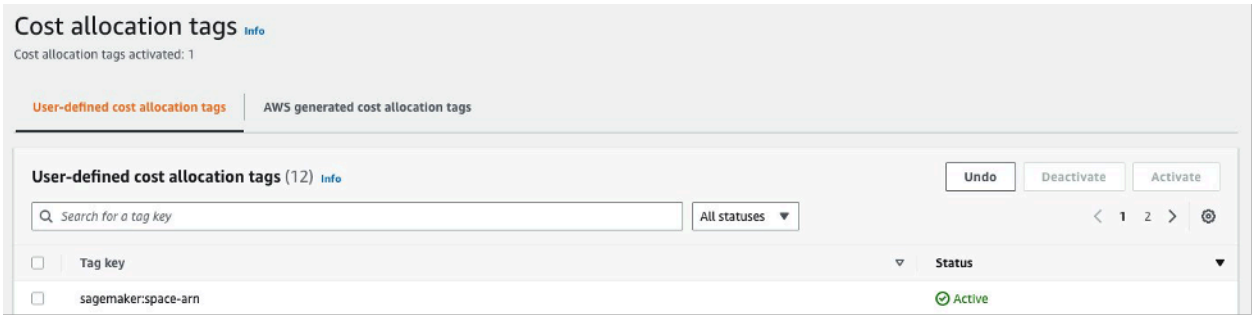
SageMaker Studio kennzeichnet neue SageMaker Ressourcen wie Trainingsjobs, Verarbeitungsjobs und Kernel-Apps jetzt automatisch mit den jeweiligen `sagemaker:domain-arn` Ressourcen. Auf einer detaillierteren Ebene kennzeichnet die Ressource SageMaker außerdem mit dem `sagemaker:user-profile-arn` Oder,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

SageMaker Domänen-EFS-Volumes sind mit einem Schlüssel gekennzeichnet, der `ManagedByAmazonSageMakerResource` mit dem Wert des Domänen-ARN benannt ist. Sie verfügen nicht über detaillierte Tags, mit denen sich die Speicherplatznutzung auf Benutzerebene nachvollziehen lässt. Administratoren können das EFS-Volume jedoch für eine maßgeschneiderte Überwachung an eine EC2-Instance anhängen.

Kostenüberwachung

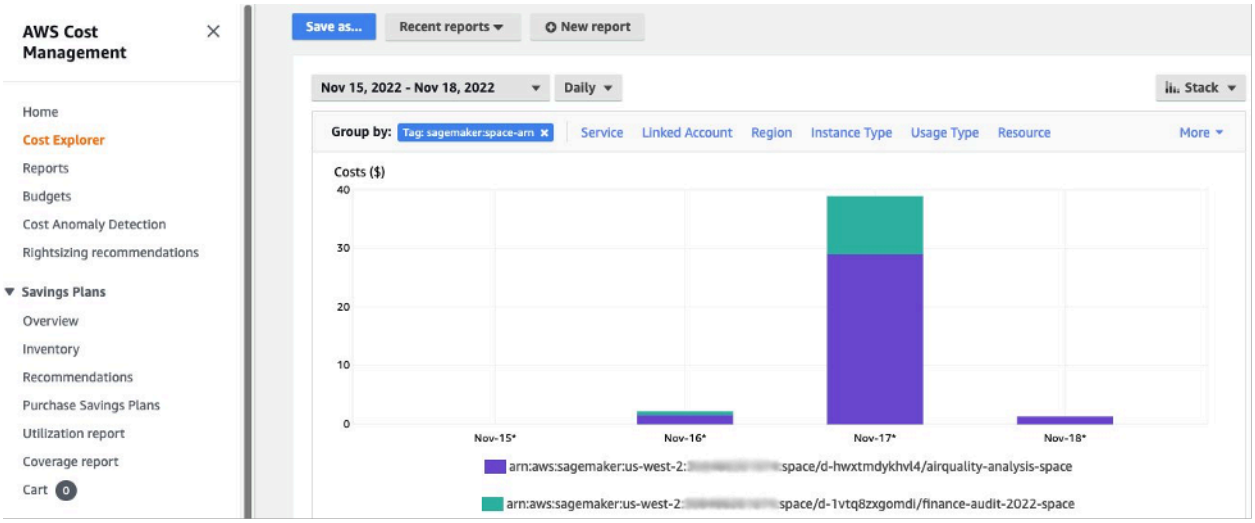
Automatisierte Tags ermöglichen es Administratoren, Ihre ML-Ausgaben mithilfe von out-of-the-box Lösungen wie und sowie benutzerdefinierten Lösungen AWS Budgets, die auf den Daten aus AWSKosten AWS Cost Explorer- und Nutzungsberichten (CURs) basieren, nachzuverfolgen, zu melden und zu überwachen.

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



Space ARN als Kostenzuweisungs-Tags im Cost Explorer aktiviert

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

Kontrolle der Kosten

Wenn der erste SageMaker Studio-Benutzer eingebunden ist, wird ein EFS-Volume für die Domain SageMaker erstellt. Für dieses EFS-Volume 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 SageMaker Amazon-Preisübersicht.

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

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.

Personalisierung

Lebenszyklus-Konfiguration

Lebenszykluskonfigurationen sind Shell-Skripts, die durch SageMaker Studio-Lebenszyklusereignisse initiiert werden, z. B. durch das Starten eines neuen Studio-Notebooks. SageMaker Sie können diese Shell-Skripts verwenden, um Anpassungen für Ihre SageMaker 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: Amazon SageMaker Studio mithilfe von Lebenszykluskonfigurationen anpassen.

Benutzerdefinierte Bilder für SageMaker Studio-Notebooks

Studio-Notebooks werden mit einer Reihe von vorgefertigten Images geliefert, die aus dem Amazon SageMaker Python SDK und der neuesten Version der IPython-Laufzeit oder des IPython-Kernels bestehen. Mit dieser Funktion können Sie Ihre eigenen benutzerdefinierten Bilder auf Amazon-Notebooks übertragen. SageMaker 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 neueste Versionen beliebter ML-Frameworks wie TensorFlow MXNet oder andere. PyTorch
- Integrieren Sie benutzerdefinierten Code oder lokal entwickelte Algorithmen in SageMaker 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.

Eine ausführliche Anleitung zum Erstellen eines benutzerdefinierten Images finden Sie unter Benutzerdefiniertes SageMaker Image erstellen.

JupyterLab Erweiterungen

Mit SageMaker 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 von selbst in den SageMaker 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 deutlich einfacher. Sie können die oben genannten Erweiterungen über Bash-Skripte installieren. Öffnen Sie in SageMaker Studio beispielsweise 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 Studio ist mit einer Jupyter-Git-Erweiterung vorinstalliert, mit der Benutzer eine maßgeschneiderte URL eines Git-Repositorys eingeben, sie in Ihr EFS-Verzeichnis klonen, Änderungen per Push 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-date Anweisungen 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 Studio-Notebooks verwenden Amazon EFS als persistente Speicherebene. 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 EFS unterstützt und sind zwischen Kernel-, App- oder Studio-Neustarts persistent. 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
```

Ausführliche Anweisungen 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 Studio-Plattform zu ermöglichen.

Anhang

Vergleich von Mehrmandantenfähigkeit

Tabelle 2 – Vergleich mit mehreren Mandanten

Multi-Domain	Mehrere Konten	Attributbasierte Zugriffsk ontrolle (ABAC) innerhalb einer einzelnen Domain
Die Ressourcenisolierung wird mithilfe von Tags erreicht. SageMaker Studio markiert automatisch alle Ressourcen mit dem Domänen-ARN und dem Benutzerprofil-/Bereichs-ARN.	Jeder Mandant befindet sich in seinem eigenen Konto, sodass es eine absolute Ressource nisolierung gibt.	Die Ressourcenisolierung wird mithilfe von Tags erreicht. Benutzer müssen das Tagging von erstellten Ressourcen für ABAC verwalten.
Listen-APIs können nicht durch Tags eingeschränkt werden. Die Benutzero berflächenfilterung von Ressourcen erfolgt für gemeinsam genutzte Bereiche. API-Aufrufe auflisten , die über die AWS CLI oder das Boto3-SDK ausgeführ t werden, listet jedoch Ressourcen in der gesamten Region auf.	Die Isolation von Listen-AP Is ist auch möglich, da sich Mandanten in ihren dediziert en Konten befinden.	Listen-APIs können nicht durch Tags eingeschränkt werden. Auflisten von API-Aufrufen, die über die AWS CLI oder das Boto3-SDK erfolgen, listet Ressourcen in der gesamten Region auf.
SageMaker Die Studio-Re chen- und Speicherkosten pro Mandant können leicht überwacht werden, indem der	SageMaker Die Studio-Re chen- und Speicherkosten pro Mandant lassen sich einfach mit einem dedizierten Konto überwachen.	SageMaker Die Studio-Re chenkosten pro Mandant müssen mit benutzerdefinierte n Tags berechnet werden.

Multi-Domain	Mehrere Konten	Attributbasierte Zugriffsk ontrolle (ABAC) innerhalb einer einzelnen Domain
Domain-ARN als Kostenzuo rdnungs-Tag verwendet wird.		SageMaker Die Studio-Sp eicherkosten können pro Domain nicht überwacht werden, da alle Mandanten dasselbe EFS-Volume verwenden.
Service Quotas werden auf Kontoebene festgelegt, sodass ein einzelner Mandant weiterhin alle Ressourcen verbrauchen kann.	Servicekontingente können für jeden Mandanten auf Kontoebene festgelegt werden.	Service Quotas werden auf Kontoebene festgelegt, sodass ein einzelner Mandant weiterhin alle Ressourcen verbrauchen kann.
Die Skalierung auf mehrere Mandanten kann über Infrastru cture as Code (IaC) oder Service Catalog erfolgen.	Die Skalierung auf mehrere Mandanten umfasst Organisat ionen und den Verkauf mehrerer Konten.	Die Skalierung benötigt für jeden neuen Mandanten eine mandantenspezifische Rolle, und Benutzerprofile müssen manuell mit Mandantennamen gekennzeichnet werden.
Die Zusammenarbeit zwischen Benutzern innerhalb eines Mandanten ist durch gemeinsam genutzte Bereiche möglich.	Die Zusammenarbeit zwischen Benutzern innerhalb eines Mandanten ist durch gemeinsam genutzte Bereiche möglich.	Alle Mandanten haben Zugriff auf denselben gemeinsam genutzten Bereich für die Zusammenarbeit.

SageMaker Studio-Domain-Backup und -Wiederherstellung

Folgen Sie im Falle eines versehentlichen EFS-Löschvorgangs oder wenn eine Domain aufgrund von Änderungen des Netzwerks oder der Authentifizierung neu erstellt werden muss, diesen Anweisungen.

Option 1: Sichern von vorhandenen EFS mit EC2

SageMaker Studio-Domain-Backup

1. Auflisten von Benutzerprofilen und Leerzeichen in SageMaker Studio (CLI ,SDK).
2. Ordnen Sie Benutzerprofile/-räume UIDs auf EFS zu.
   a. Beschreiben Sie für jeden Benutzer in der Liste der Benutzer/Bereiche das Benutzerprofil/den Arbeitsbereich (CLI ,SDK).
   b. Ordnen Sie Benutzerprofil/-raum zu zu`HomeEfsFileSystemUid`.
   c. Ordnen Sie das Benutzerprofil zu, `UserSettings['ExecutionRole']` wenn Benutzer unterschiedliche Ausführungsrollen haben.
   d. Identifizieren Sie die standardmäßige Space-Ausführungsrolle.
3. Erstellen Sie eine neue Domäne und geben Sie die Standardrolle für die Ausführung von Bereichen an.
4. Erstellen Sie Benutzerprofile und Leerzeichen.
   - Erstellen Sie für jeden Benutzer in der Benutzerliste ein Benutzerprofil (CLI ,SDK) mithilfe der Zuweisung der Ausführungsrolle.
5. Erstellen Sie eine Zuordnung für die neuen EFS- und UIDs .
   a. Beschreiben Sie für jeden Benutzer in der Benutzerliste das Benutzerprofil (CLI ,SDK).
   b. Ordnen Sie das Benutzerprofil zu zu`HomeEfsFileSystemUid`.
6. Löschen Sie optional alle Apps, Benutzerprofile und Leerzeichen und löschen Sie dann die Domain.

EFS-Sicherung

Verwenden Sie die folgenden Anweisungen, um EFS zu sichern:

1. Starten Sie die EC2-Instance und fügen Sie die eingehenden/ausgehenden Sicherheitsgruppen der alten SageMaker Studio-Domain an die neue EC2-Instance an (lassen Sie NFS-Datenverkehr über TCP auf Port 2049 zu. Weitere Informationen finden Sie unter Studio SageMaker -Notebooks in einer VPC mit externen Ressourcen verbinden.
2. Mounten Sie das SageMaker Studio-EFS-Volume auf der neuen EC2-Instance. Weitere Informationen finden Sie unter Mounting von EFS-Dateisystemen.

3. Kopieren Sie die Dateien in den lokalen EBS-Speicher: `>sudo cp -rp /efs /studio-backup:`
   a. Fügen Sie die neuen Domänensicherheitsgruppen an die EC2-Instance an.
   b. Mounten Sie das neue EFS-Volume an die EC2-Instance.
   c. Kopieren Sie Dateien auf das neue EFS-Volume.
   d. Für jeden Benutzer in der Benutzersammlung:
      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: Sichern von vorhandenen EFS mithilfe von S3 und Lebenszykluskonfiguration

1. Weitere Informationen finden Sie unter Migrieren Ihrer Arbeit zu einer Amazon- SageMaker Notebook-Instance mit Amazon Linux 2.
2. Erstellen Sie einen S3-Bucket für die Sicherung (z. B. `>studio-backup`.
3. Listen Sie alle Benutzerprofile mit Ausführungsrollen auf.
4. Legen Sie in der aktuellen SageMaker Studio-Domäne ein Standard-LCC-Skript auf Domänenebene fest.
   - Kopieren Sie im LCC alles in `/home/sagemaker-user` das Benutzerprofilpräfix in S3 (z. B. `s3://studio-backup/studio-user1`).
5. Starten Sie alle standardmäßigen Jupyter Server-Apps neu (für die Ausführung des LCC).
6. Löschen Sie alle Apps, Benutzerprofile und Domänen.
7. Erstellen Sie eine neue SageMaker Studio-Domain.
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 LCC-Konfiguration (CLI ,SDK).

SageMaker Studio-Zugriff mit SAML-Assertion

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 Funktions-URL oder 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 annehmen können, um das API Gateway aufzurufen. Diese Rolle sollte in SAML-Assertion als Attribut im folgenden Format übergeben werden:
   - Attributname: https://aws.amazon.com/SAML/Attributes/Role
   - Attributwert: `<IdentityProviderARN>`, `<RoleARN>`
6. Aktualisieren Sie den SAML Assertion Consumer Service (ACS)-Endpunkt auf die `SAMLValidator` Aufruf-URL.

Beispielcode für die SAML-Validierung:

```
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 (AWSBlog)
- Konfiguration von Amazon SageMaker Studio für Teams und Gruppen mit vollständiger Ressourcenisolierung (AWSBlog)
- Einführung in Amazon SageMaker Studio mit AWS SSO und Okta Universal Directory (Blog) AWS
- So konfigurieren Sie SAML 2.0 für den AWS Kontoverbund (Okta-Dokumentation)
- Erstellen Sie eine sichere Plattform für Machine Learning für Unternehmen auf AWS (AWStechnischer Leitfaden)
- Passen Sie Amazon SageMaker Studio mithilfe von Lebenszykluskonfigurationen an (AWSBlog)
- Bringen Sie Ihr eigenes benutzerdefiniertes Container-Image in Amazon SageMaker Studio-Notizbücher (AWSBlog)
- Erstellen Sie benutzerdefinierte SageMaker Projektvorlagen — Best Practices (AWSBlog)
- Implementierung eines Modells mit mehreren Konten mit Amazon SageMaker Pipelines (Blog) AWS
- Teil 1: Wie NatWest Group eine skalierbare, sichere und nachhaltige MLOps-Plattform aufgebaut hat (Blog) AWS
- Sichere vorsignierte URLs von Amazon SageMaker Studio Teil 1: Grundlegende Infrastruktur (Blog) AWS

Beitragende Faktoren

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 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.