AWS Livro branco

SageMaker Práticas recomendadas de administração do Studio

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

SageMaker Práticas recomendadas de administração do Studio: AWS Livro branco

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

As marcas comerciais e imagens comerciais da Amazon não podem ser usadas no contexto de nenhum produto ou serviço que não seja da Amazon, nem de qualquer maneira que possa gerar confusão entre os clientes ou que deprecie ou desprestigie a Amazon. Todas as outras marcas comerciais que não são propriedade da Amazon pertencem aos respectivos proprietários, os quais podem ou não ser afiliados, estar conectados ou ser patrocinados pela Amazon.

Table of Contents

SageMaker Práticas recomendadas de administração do Studio

Data de publicação: 25 de abril de 2023 (Revisões do documento)

Resumo

O Amazon SageMaker AI Studio fornece uma interface visual única, baseada na web, na qual você pode realizar todas as etapas de desenvolvimento de aprendizado de máquina (ML), o que melhora a produtividade da equipe de ciência de dados. SageMaker O AI Studio oferece acesso, controle e visibilidade completos de cada etapa necessária para criar, treinar e avaliar modelos.

Neste whitepaper, discutimos as melhores práticas para assuntos que incluem modelo operacional, gerenciamento de domínio, gerenciamento de identidades, gerenciamento de permissões, gerenciamento de rede, log, monitoramento e personalização. As melhores práticas discutidas aqui se destinam à implantação do SageMaker AI Studio corporativo, incluindo implantações multilocatárias. Este documento é destinado a administradores de plataformas de ML, engenheiros de ML e arquitetos de ML.

Você é Well-Architected?

O Well-Architected Framework da AWS ajuda você a entender os prós e os contras das decisões que você toma ao criar sistemas na nuvem. Os seis pilares do framework permitem a você conhecer as melhores práticas de arquitetura para criar e operar sistemas confiáveis, seguros, econômicos e sustentáveis na nuvem. Usando o AWS Well-Architected Tool, disponível gratuitamente no AWS Management Console, você pode analisar suas workloads em relação a essas práticas recomendadas respondendo a um conjunto de perguntas para cada pilar.

No Machine Learning Lens, nos concentramos em como projetar, implantar e arquitetar suas cargas de trabalho de machine learning no Nuvem AWS. Essa lente complementa as práticas recomendadas descritas no Well-Architected Framework.

Introdução

Ao administrar o SageMaker AI Studio como sua plataforma de ML, você precisa de orientação sobre as melhores práticas para tomar decisões informadas para ajudá-lo a escalar sua plataforma de

ML à medida que suas cargas de trabalho crescem. Para provisionar, operacionalizar e escalar sua plataforma de ML, considere o seguinte:

- Escolha o modelo operacional certo e organize seus ambientes de ML para atender aos seus objetivos de negócios.
- Escolha como configurar a autenticação de domínio do SageMaker AI Studio para identidades de usuário e considere as limitações em nível de domínio.
- Decida como federar a identidade e a autorização de seus usuários na plataforma de ML para controles de acesso e auditoria refinados.
- Considere configurar permissões e barreiras de proteção para várias perfis de suas personas de ML.
- Planeje sua topologia de rede de nuvem privada virtual (VPC), considerando a sensibilidade da carga de trabalho de ML, o número de usuários, os tipos de instância, os aplicativos e os trabalhos lançados.
- Classifique e proteja seus dados em repouso e em trânsito com criptografia.
- Considere como registrar e monitorar várias interfaces de programação de aplicativos (APIs) e atividades do usuário para fins de conformidade.
- Personalize a experiência do notebook SageMaker AI Studio com suas próprias imagens e scripts de configuração do ciclo de vida.

Modelo operacional

Um modelo operacional é uma estrutura que reúne pessoas, processos e tecnologias para ajudar uma organização a oferecer valor comercial de maneira escalável, consistente e eficiente. O modelo operacional de ML fornece um processo padrão de desenvolvimento de produtos para equipes em toda a organização. Há três modelos para implementar o modelo operacional, dependendo do tamanho, da complexidade e dos fatores de negócios:

- Equipe centralizada de ciência de dados — Nesse modelo, todas as atividades de ciência de dados são centralizadas em uma única equipe ou organização. Isso é semelhante ao modelo Center of Excellence (COE), em que todas as unidades de negócios recorrem a essa equipe para projetos de ciência de dados.
- Equipes descentralizadas de ciência de dados — Nesse modelo, as atividades de ciência de dados são distribuídas em diferentes perfis ou divisões de negócios, ou com base em diferentes linhas de produtos.
- Equipes federadas de ciência de dados — Nesse modelo, perfis de serviços compartilhados, como repositórios de código, pipelines de integração contínua e entrega contínua (CI/CD), etc., são gerenciadas pela equipe centralizada, e cada unidade de negócios ou perfil de nível de produto é gerenciada por equipes descentralizadas. Isso é semelhante ao modelo hub and spoke, em que cada unidade de negócios tem suas próprias equipes de ciência de dados; no entanto, essas equipes de unidades de negócios coordenam suas atividades com a equipe centralizada.

Antes de decidir lançar seu primeiro domínio de estúdio para casos de uso de produção, considere seu modelo operacional e as AWS melhores práticas para organizar seu ambiente. Para obter mais informações, consulte Organizando seu AWS ambiente usando várias contas.

A próxima seção fornece orientação sobre como organizar sua estrutura de conta para cada um dos modelos operacionais.

Estrutura de conta recomendada

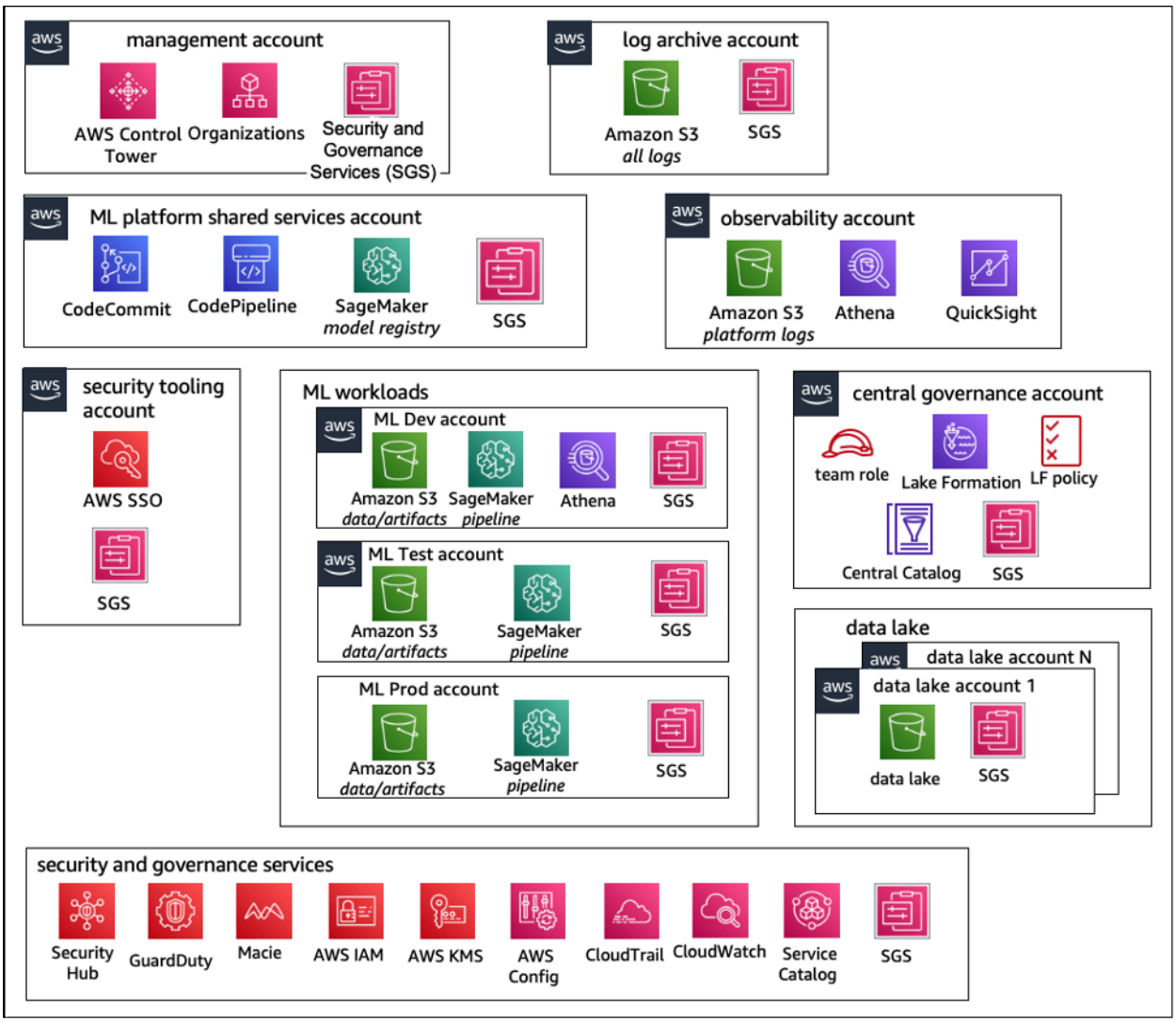
Nesta seção, apresentamos brevemente uma estrutura de conta do modelo operacional com a qual você pode começar e modificar de acordo com os requisitos operacionais da sua organização. Independentemente do modelo operacional escolhido, recomendamos implementar as seguintes práticas recomendadas comuns:

- Use AWS Control Tower para configuração, gerenciamento e governança de suas contas.
- Centralize suas identidades com seu Provedor de Identidade (IdP) e a Central de AWS IAM Identidades com uma conta delegada do Security Tooling de administrador e permita o acesso seguro às cargas de trabalho.
- Execute cargas de trabalho de ML com isolamento em nível de conta em cargas de trabalho de desenvolvimento, teste e produção.
- Transmita registros de carga de trabalho de ML para uma conta de arquivamento de logs e, em seguida, filtre e aplique a análise de registros em uma conta de observabilidade.
- Execute uma conta de governança centralizada para provisionar, controlar e auditar o acesso aos dados.
- Incorpore serviços de segurança e governança (SGS) com proteções preventivas e de deteção apropriadas em cada conta para garantir a segurança e a conformidade, de acordo com os requisitos de sua organização e carga de trabalho.

Modelo centralizado de estrutura de conta

Nesse modelo, a equipe da plataforma ML é responsável por fornecer:

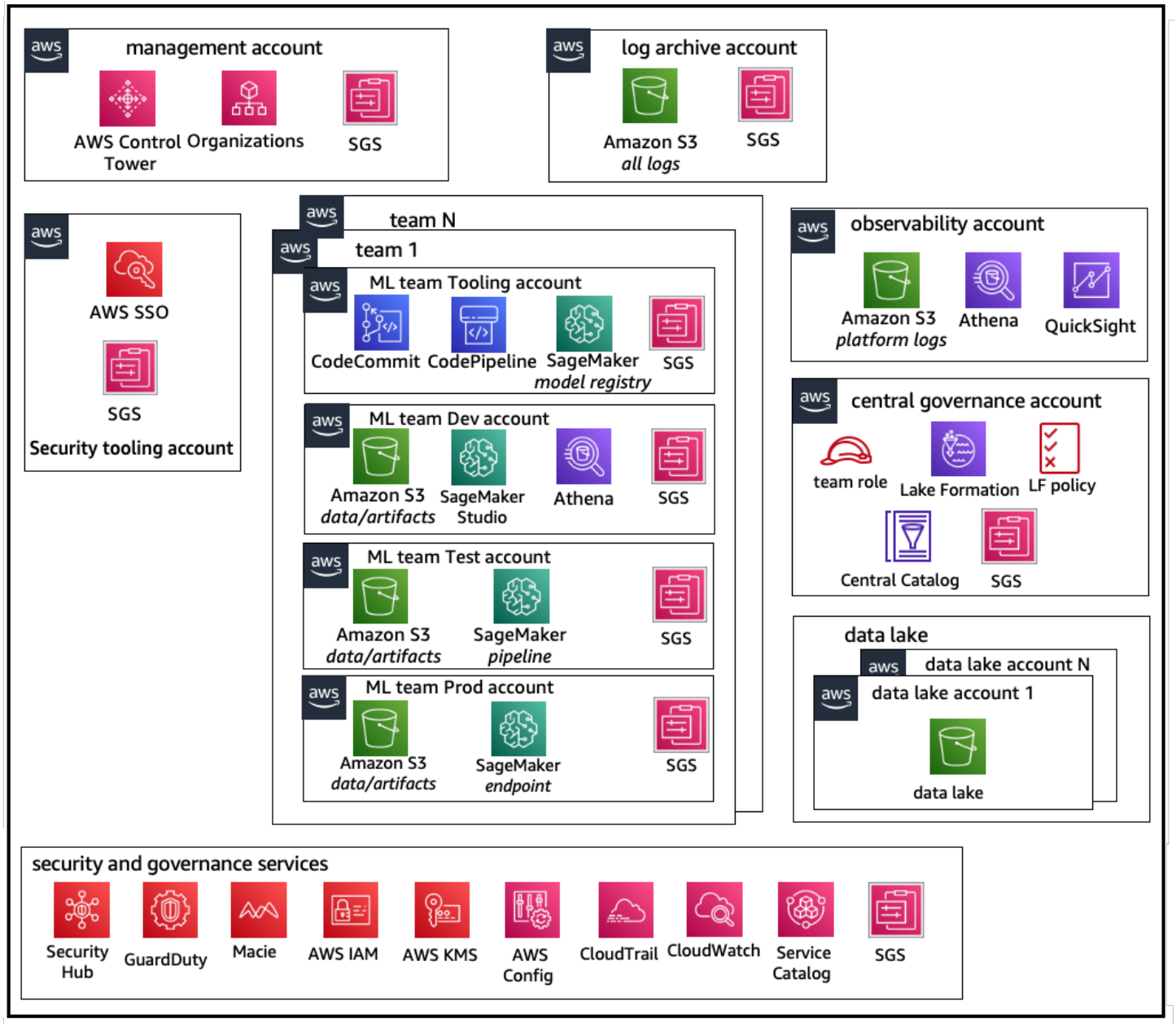
- Uma conta de ferramentas de serviços compartilhados que atende aos requisitos de Machine Learning Operations (MLOps) em todas as equipes de ciência de dados.
- Contas de desenvolvimento, teste e produção de cargas de trabalho de ML que são compartilhadas entre as equipes de ciência de dados.
- Políticas de governança para garantir que a carga de trabalho de cada equipe de ciência de dados seja executada isoladamente.
- Práticas recomendadas comuns.



Estrutura de contas do modelo operacional centralizado

Estrutura de conta modelo descentralizada

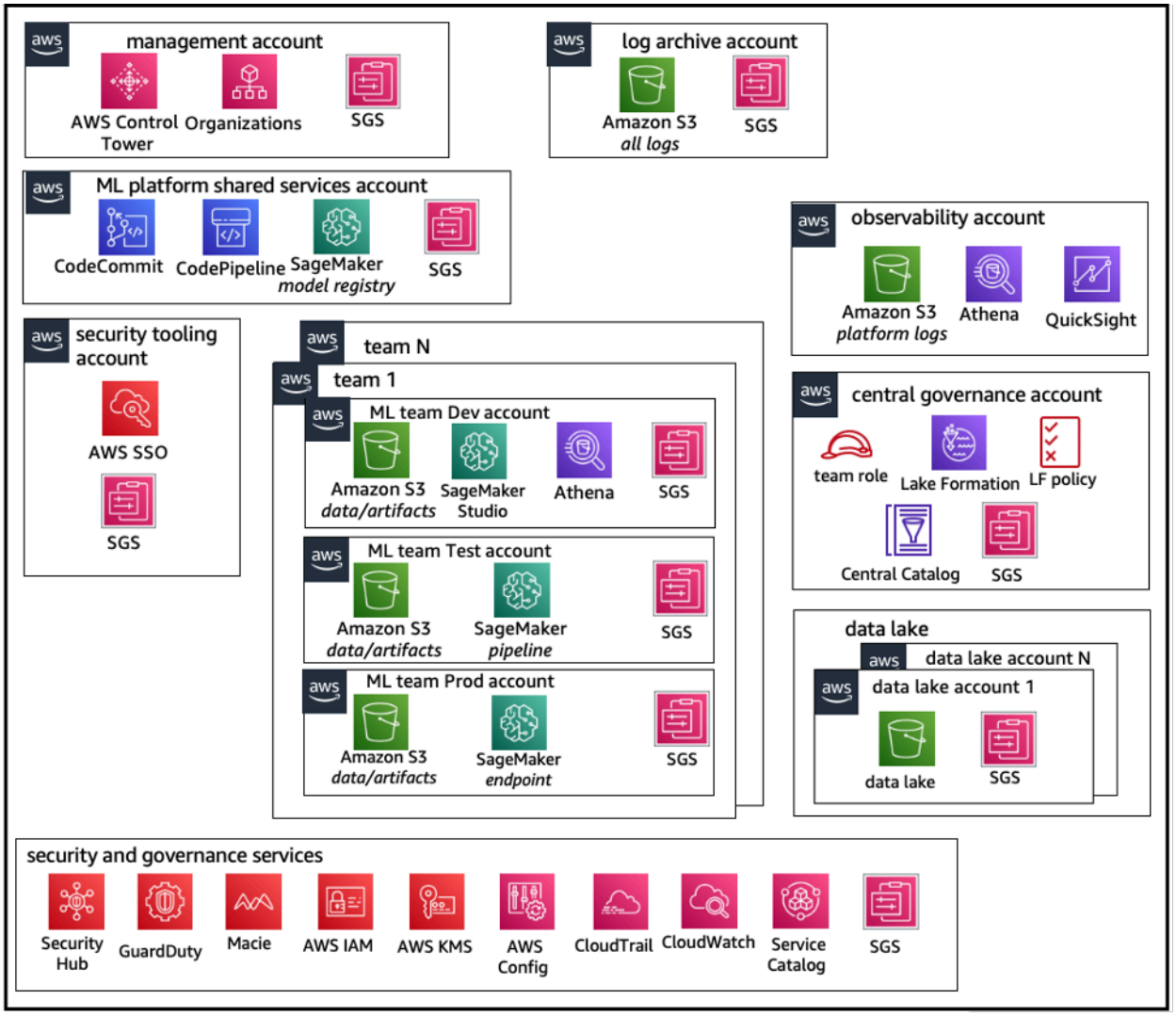
Nesse modelo, cada equipe de ML opera de forma independente para provisionar, gerenciar e governar contas e recursos de ML. No entanto, recomendamos que as equipes de ML usem uma abordagem centralizada de observabilidade e modelo de governança de dados para simplificar a governança de dados e o gerenciamento de auditoria.



Estrutura de contas do modelo operacional descentralizado

Estrutura de conta do modelo federado

Esse modelo é semelhante ao modelo centralizado; no entanto, a principal diferença é que cada science/ML team gets their own set of development/test/production carga de trabalho de dados permite um isolamento físico robusto de seus recursos de ML e também permite que cada equipe escale de forma independente sem afetar outras equipes.



Estrutura de contas do modelo operacional federado

Multilocação da plataforma ML

A multilocação é uma arquitetura de software em que uma única instância de software pode atender a vários grupos de usuários distintos. Um locatário é um grupo de usuários que compartilham acesso comum com privilégios específicos à instância do software. Por exemplo, se você estiver criando vários produtos de ML, cada equipe de produto com requisitos de acesso semelhantes pode ser considerada locatária ou equipe.

Embora seja possível implementar várias equipes em uma instância do SageMaker AI Studio (como o SageMaker AI Domain), avalie essas vantagens em relação a compensações, como raio de explosão, atribuição de custos e limites de nível de conta, ao reunir várias equipes em um único

domínio do AI Studio. SageMaker Saiba mais sobre essas compensações e as melhores práticas nas seções a seguir.

Se você precisar de isolamento absoluto de recursos, considere implementar domínios do SageMaker AI Studio para cada inquilino em uma conta diferente. Dependendo dos seus requisitos de isolamento, você pode implementar várias linhas de negócios (LOBs) como vários domínios em uma única conta e região. Use espaços compartilhados para colaboração quase em tempo real entre membros da mesma equipe/LOB. Com vários domínios, você ainda usará políticas e permissões de gerenciamento de acesso à identidade (IAM) para garantir o isolamento dos recursos.

SageMaker Os recursos de IA criados a partir de um domínio são marcados automaticamente com o domínio Amazon Resource Name (ARN) e o perfil ou espaço do usuário ARN para facilitar o isolamento dos recursos. Para exemplos de políticas, consulte a documentação de isolamento de recursos de domínio. Lá, você pode ver a referência detalhada de quando usar uma estratégia de várias contas ou vários domínios, junto com as comparações de recursos na documentação, e você pode ver exemplos de scripts para preencher as tags dos domínios existentes no repositório. GitHub

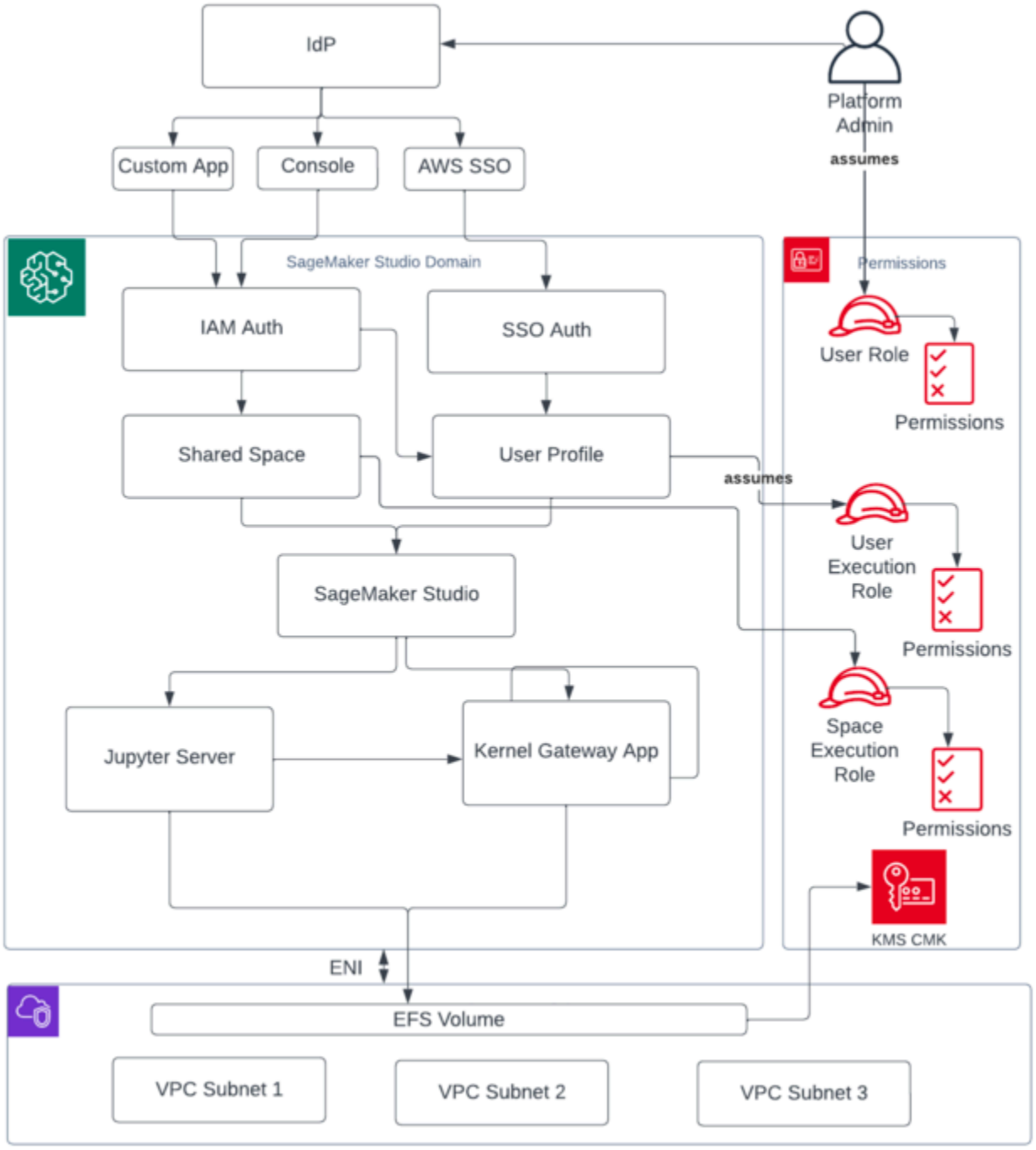
Por fim, você pode implementar uma implantação de autoatendimento dos recursos do SageMaker AI Studio em várias contas usando o. AWS Service Catalog Para obter mais informações, consulte Gerenciar AWS Service Catalog produtos em vários Contas da AWSRegiões da AWS e.

Gerenciamento de domínio

Um domínio Amazon SageMaker AI consiste em:

- Um volume associado do Amazon Elastic File System (AmazonEFS)
- Uma lista de usuários autorizados
- Uma variedade de configurações de segurança, aplicativos, políticas e Amazon Virtual Private Cloud (AmazonVPC)

O diagrama a seguir fornece uma visão de alto nível dos vários componentes que constituem um SageMaker AIStudio domínio:

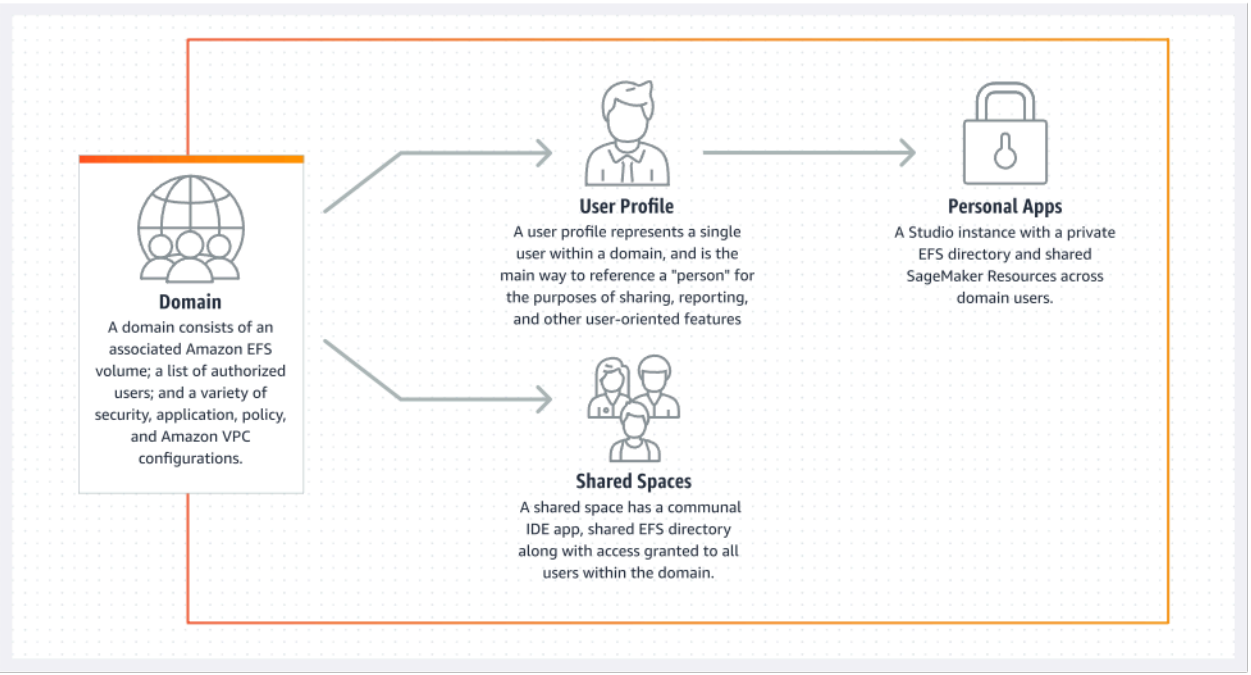


Visão de alto nível de vários componentes que constituem um domínio do SageMaker AI Studio

Vários domínios e espaços compartilhados

O Amazon SageMaker AI agora suporta a criação de vários domínios de SageMaker IA em um único Região da AWS para cada conta. Cada domínio pode ter suas próprias configurações de domínio, como modo de autenticação, e configurações de rede, como VPC sub-redes. Um perfil de usuário não pode ser compartilhado entre domínios. Se um usuário humano fizer parte de várias equipes separadas por domínios, crie um perfil de usuário para o usuário em cada domínio. Consulte a Visão geral de vários domínios para saber mais sobre o preenchimento de tags para domínios existentes.

Cada domínio configurado no modo de IAM autenticação pode usar o espaço compartilhado para colaboração quase em tempo real entre os usuários. Com um espaço compartilhado, os usuários têm acesso a um EFS diretório compartilhado da Amazon e a um JupyterServeraplicativo compartilhado para a interface do usuário e podem coeditar quase em tempo real. A marcação automática de recursos criados por espaços compartilhados permite que os administradores acompanhem os custos em um nível de projeto. A JupyterServer interface compartilhada também filtra recursos, como experimentos e entradas de registro de modelos, para que somente itens relevantes para o esforço compartilhado de ML sejam exibidos. O diagrama a seguir fornece uma visão geral dos aplicativos privados e espaços compartilhados em cada domínio.



Visão geral de aplicativos privados e espaços compartilhados em um único domínio

Configure espaços compartilhados em seu domínio

Os espaços compartilhados geralmente são criados para um determinado empreendimento ou projeto de ML em que os membros de um único domínio precisam de acesso quase em tempo real ao mesmo armazenamento de arquivos subjacente e. IDE O usuário pode acessar, ler, editar e compartilhar seus cadernos quase em tempo real, o que lhe dá o caminho mais rápido para começar a iterar com seus colegas.

Para criar um espaço compartilhado, você deve primeiro designar um perfil de execução padrão do espaço que governará as permissões de qualquer usuário que utilize o espaço. No momento da redação deste artigo, todos os usuários em um domínio terão acesso a todos os espaços compartilhados em seu domínio. Consulte Criar um espaço compartilhado para obter a documentação mais recente sobre como adicionar espaços compartilhados a um domínio existente.

Configure seu domínio para IAM federação

Antes de configurar a federação AWS Identity and Access Management (IAM) para seu domínio do SageMaker AI Studio, você precisa configurar uma função de usuário IAM da federação (como administrador da plataforma) no seu IdP, conforme discutido na seção Gerenciamento de identidade.

Para obter instruções detalhadas sobre como configurar o SageMaker AI Studio com a IAM opção, consulte Integração ao SageMaker domínio da Amazon usando o IAM Identity Center.

Configure seu domínio para federação de login único () SSO

Para usar a federação de login único (SSO), você precisa habilitar AWS IAM Identity Center sua conta AWS Organizationsde gerenciamento na mesma região em que precisa executar o SageMaker AI Studio. As etapas de configuração do domínio são semelhantes às etapas de IAM federação, exceto que você seleciona AWS IAM Identity Center(iDC) na seção Autenticação.

Para obter instruções detalhadas, consulte Integrar o SageMaker domínio da Amazon usando o IAM Identity Center.

SageMaker Perfil de usuário do AI Studio

Um perfil de usuário representa um único usuário dentro de um domínio e é a principal maneira de referenciar uma “pessoa” para fins de compartilhamento, relatórios e outros recursos orientados

para o usuário. Essa entidade é criada quando um usuário integra o toSageMaker AI Studio. Se um administrador convidar uma pessoa por e-mail ou importá-la do IdC, um perfil de usuário será criado automaticamente. Um perfil de usuário é o principal detentor das configurações de um usuário individual e tem uma referência ao diretório inicial privado do Amazon Elastic File System (AmazonEFS) do usuário. Recomendamos criar um perfil de usuário para cada usuário físico do aplicativo SageMaker AI Studio. Cada usuário tem seu próprio diretório dedicado na AmazonEFS, e os perfis de usuário não podem ser compartilhados entre domínios na mesma conta.

Cada perfil de usuário que compartilha o domínio do SageMaker AI Studio recebe recursos computacionais dedicados (como instâncias de SageMaker AI Amazon Elastic Compute Cloud (AmazonEC2)) para executar notebooks. As instâncias de computação alocadas para o usuário um são completamente isoladas daquelas alocadas para o usuário dois. Da mesma forma, os recursos computacionais alocados aos usuários em uma conta da AWS são completamente separados daqueles alocados aos usuários em outra conta. Cada usuário pode executar até quatro aplicativos (aplicativos) em contêineres isolados do Docker ou imagens no mesmo tipo de instância.

Aplicativo Jupyter Server

Quando você inicia um notebook Amazon SageMaker AI Studio para um usuário acessando o pré-assinado URL ou fazendo login usando o AWS IAM iDC, o aplicativo Jupyter Server é lançado na instância gerenciada pelo SageMaker serviço de IA. VPC Cada usuário obtém seu próprio aplicativo Jupyter Server dedicado em um aplicativo privado. Por padrão, o aplicativo Jupyter Server para notebooks SageMaker AI Studio é executado em uma `ml.t3.medium` instância dedicada (reservada como um tipo de instância do sistema). A computação dessa instância não é cobrada do cliente.

O aplicativo Jupyter Kernel Gateway

O aplicativo Kernel Gateway pode ser criado por meio da interface API ou do SageMaker AI Studio e é executado no tipo de instância escolhido. Esse aplicativo pode ser executado usando uma das imagens integradas do SageMaker AI Studio que são pré-configuradas com pacotes populares de ciência de dados e aprendizado profundo TensorFlow, como Apache MXNet e. PyTorch

Os usuários podem iniciar e executar vários kernels do notebook Jupyter, sessões de terminal e consoles interativos no mesmo Studio. SageMaker image/Kernel Gateway app. Users can also run up to four Kernel Gateway apps or images on the same physical instance—each isolated by its container/image

Para criar aplicativos adicionais, você precisa usar um tipo de instância diferente. Um perfil de usuário pode ter somente uma instância em execução, de qualquer tipo de instância. Por exemplo, um usuário pode executar um notebook simples usando a imagem de ciência de dados integrada do SageMaker AI Studio e outro notebook usando a TensorFlow imagem integrada, na mesma instância. Os usuários são cobrados pelo tempo em que a instância está em execução. Para evitar custos quando o usuário não está executando ativamente o SageMaker AI Studio, o usuário precisa desligar a instância. Para obter mais informações, consulte Desligar e atualizar os aplicativos do Studio.

Toda vez que você desliga e reabre um aplicativo Kernel Gateway a partir da interface do SageMaker AI Studio, esse aplicativo é iniciado em uma nova instância. Isso significa que a instalação do pacote não persiste por meio de reinicializações do mesmo aplicativo. Da mesma forma, se um usuário alterar o tipo de instância em um notebook, os pacotes instalados e as variáveis de sessão serão perdidos. No entanto, você pode usar recursos como trazer sua própria imagem e scripts de ciclo de vida para trazer os pacotes do próprio usuário para o SageMaker AI Studio e mantê-los por meio de trocas de instância e lançamentos de novas instâncias.

Volume do Amazon Elastic File System

Quando um domínio é criado, um único volume do Amazon Elastic File System (AmazonEFS) é criado para ser usado por todos os usuários dentro do domínio. Cada perfil de usuário recebe um diretório pessoal privado dentro do EFS volume da Amazon para armazenar os notebooks, GitHub repositórios e arquivos de dados do usuário. Cada espaço dentro de um domínio recebe um diretório privado dentro do EFS volume da Amazon que pode ser acessado por vários perfis de usuário. O acesso às pastas é segregado por usuário, por meio de permissões do sistema de arquivos. SageMaker O AI Studio cria uma ID de usuário global exclusiva para cada perfil ou espaço de usuário e a aplica como uma interface de sistema operacional portátil (POSIX) a user/group ID for the user's home directory on EFS, which prevents other users/spaces partir do acesso aos dados.

Backup e recuperação

Um EFS volume existente não pode ser anexado a um novo domínio de SageMaker IA. Em uma configuração de produção, certifique-se de que o EFS volume da Amazon tenha sido copiado (para outro EFS volume ou para o Amazon Simple Storage Service (Amazon S3)). Se um EFS volume for excluído acidentalmente, o administrador precisará remover e recriar o domínio do SageMaker AI Studio. O processo é o seguinte:

Faça backup da lista de perfis de usuário, espaços e do EFS usuário associado IDs (UIDs) por meio das `DescribeSpace` API chamadas `ListUserProfiles` `DescribeUserProfileList Spaces`,, e.

1. Crie um novo domínio do SageMaker AI Studio.
2. Crie os perfis e espaços do usuário.
3. Para cada perfil de usuário, copie os arquivos do backup no EFS /Amazon S3.
4. Opcionalmente, exclua todos os aplicativos e perfis de usuário no antigo domínio do SageMaker AI Studio.

Para obter instruções detalhadas, consulte a seção do apêndice Backup e recuperação de domínios do SageMaker AI Studio.

> **Note**
>
> Isso também pode ser feito por meio do `LifecycleConfigurations` para fazer o backup de dados e do S3 sempre que um usuário inicia o aplicativo.

EBSVolume da Amazon

Um volume de armazenamento do Amazon Elastic Block Store (AmazonEBS) também é anexado a cada instância do SageMaker AI Studio Notebook. Ele é usado como o volume raiz do contêiner ou da imagem em execução na instância. Embora o EFS armazenamento da Amazon seja persistente, o EBS volume da Amazon anexado ao contêiner é temporário. Os dados armazenados localmente no EBS volume da Amazon não serão mantidos se o cliente excluir o aplicativo.

Protegendo o acesso ao pré-assinado URL

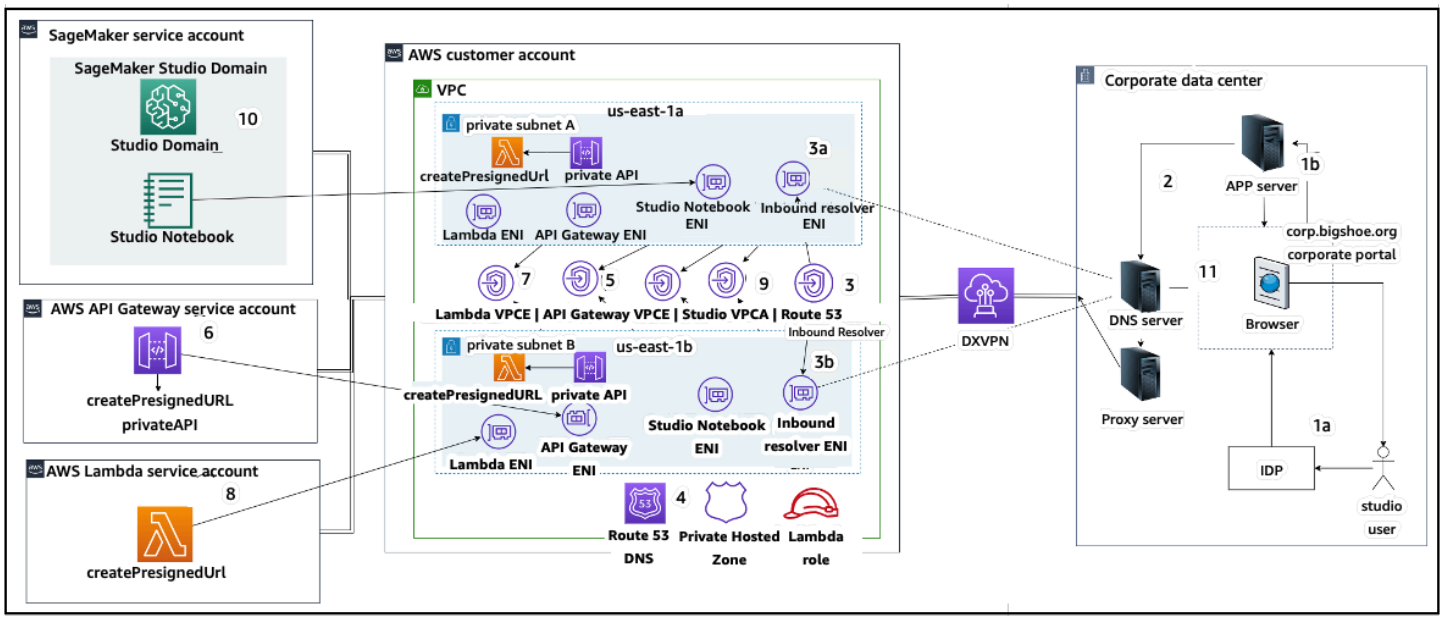
Quando um usuário do SageMaker AI Studio abre o link do notebook, o SageMaker AI Studio valida a IAM política do usuário federado para autorizar o acesso e gera e resolve o URL pré-assinado para o usuário. Como o console de SageMaker IA é executado em um domínio da Internet, esse conteúdo gerado e pré-assinado URL fica visível na sessão do navegador. Isso representa um vetor de ameaça indesejado para roubo de dados e para a obtenção de acesso aos dados do cliente quando os controles de acesso adequados não são aplicados.

O Studio oferece suporte a alguns métodos para aplicar controles de acesso contra roubo de URL dados pré-assinados:

- Validação do IP do cliente usando a condição IAM de política `aws:sourceIp`
- VPCValidação do cliente usando a IAM condição `aws:sourceVpc`
- Validação do VPC endpoint do cliente usando a condição da IAM política `aws:sourceVpce`

Quando você acessa os notebooks do SageMaker AI Studio a partir do console do SageMaker AI, a única opção disponível é usar a validação de IP do cliente com a condição `aws:sourceIp` da IAM política. No entanto, você pode usar produtos de roteamento de tráfego do navegador, como o Zscaler, para garantir a escala e a conformidade do acesso à Internet da sua força de trabalho. Esses produtos de roteamento de tráfego geram seu próprio IP de origem, cujo intervalo de IP não é controlado pelo cliente corporativo. Isso impossibilita que esses clientes corporativos usem a condição `aws:sourceIp`.

Para usar a validação do VPC endpoint do cliente usando a condição da IAM políticaaws:sourceVpce, a criação de um terminal pré-assinado URL precisa se originar no mesmo cliente VPC em que o SageMaker AI Studio está implantado, e a resolução do pré-assinado URL precisa ocorrer por meio de um endpoint do SageMaker AI Studio VPC no cliente. VPC Essa resolução do pré-assinado URL durante o tempo de acesso para usuários da rede corporativa pode ser realizada usando regras de DNS encaminhamento (tanto no Zscaler quanto no corporativoDNS) e, em seguida, no endpoint do cliente VPC usando um resolvedor de entrada do Amazon Route 53, conforme mostrado na arquitetura a seguir:



Acessando o Studio pré-assinado URL com VPC endpoint pela rede corporativa

Para step-by-step obter orientação sobre a configuração da arquitetura anterior, consulte Secure Amazon SageMaker AI Studio presigned URLs Part 1: Foundational infrastructure.

SageMaker Cotas e limites de domínio de IA

- SageMaker A SSO federação de domínios do AI Studio é suportada somente na região, em todas as contas membros da AWS organização em que o AWS Identity Center é provisionado.
- Atualmente, os espaços compartilhados não são compatíveis com domínios configurados com o AWS Identity Center.
- VPCe a configuração da sub-rede não pode ser alterada após a criação do domínio. No entanto, você pode criar um novo domínio com uma configuração diferente VPC e de sub-rede.
- O acesso ao domínio não pode ser alternado entre os SSO modos IAM e após a criação do domínio. Você pode criar um novo domínio com um modo de autenticação diferente.
- Há um limite de quatro aplicativos de gateway do kernel por tipo de instância lançado para cada usuário.
- Cada usuário pode iniciar somente uma instância de cada tipo de instância.
- Há limites nos recursos consumidos em um domínio, como o número de instâncias lançadas por tipos de instância e o número de perfis de usuário que podem ser criados. Consulte a página de cota de serviço para obter uma lista completa dos limites de serviço.
- Os clientes podem enviar um caso de suporte corporativo com justificativa comercial para aumentar os limites de recursos padrão, como número de domínios ou perfis de usuário, sujeitos a barreiras de proteção em nível de conta.
- O limite rígido do número de aplicativos simultâneos por conta é de 2.500 aplicativos. Os limites de domínios e perfis de usuário dependem desse limite rígido. Por exemplo, uma conta pode ter um único domínio com 1.000 perfis de usuário ou 20 domínios com 50 perfis de usuário cada.

Gerenciamento de identidades

Esta seção discute como os usuários da força de trabalho em um diretório corporativo se federam Contas da AWS e SageMaker acessam o AI Studio. Primeiro, descreveremos brevemente como usuários, grupos e perfis são mapeados e como funciona a federação de usuários.

Usuários, grupos e perfil

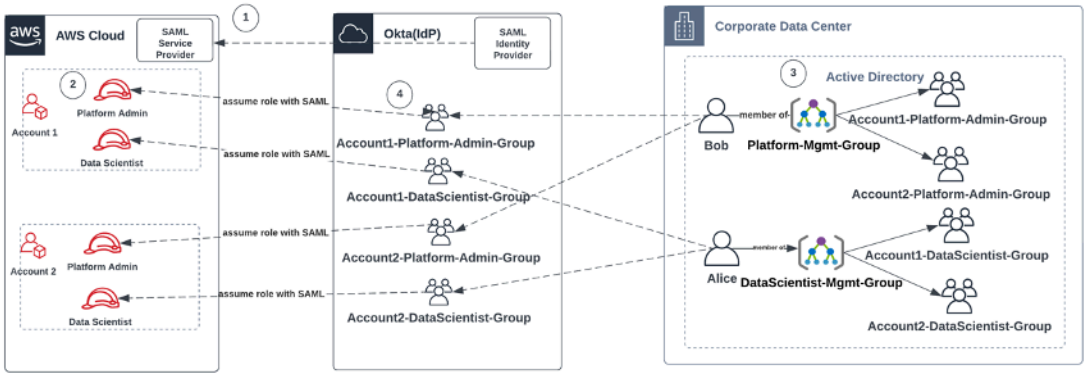
Em AWS, as permissões de recursos são gerenciadas usando usuários, grupos e funções. Os clientes podem gerenciar seus usuários e grupos por meio de ou em um diretório corporativoIAM, como o Active Directory (AD), habilitado por meio de um IdP externo, como o Okta, que permite autenticar os usuários em vários aplicativos executados na nuvem e no local.

Conforme discutido na seção Gerenciamento de Identidades do AWS Security Pillar, é uma prática recomendada gerenciar suas identidades de usuário em um IdP central, pois isso ajuda a se integrar facilmente aos seus processos de RH de back-end e ajuda a gerenciar o acesso aos usuários da sua força de trabalho.

IdPs como o Okta, permitem que os usuários finais se autentiquem em uma ou mais Contas da AWS e tenham acesso a funções específicas usando a linguagem de marcação SSO de asserção de segurança (). SAML Os administradores do IdP têm a capacidade de baixar funções do Contas da AWS IdP e atribuí-las aos usuários. Ao fazer login AWS, os usuários finais recebem uma AWS tela que exibe uma lista de AWS funções atribuídas a eles em uma ou mais Contas da AWS. Eles podem selecionar o perfil a ser assumido para o login, o que define suas permissões durante a sessão autenticada.

Um grupo deve existir no IdP para cada combinação específica de conta e perfil à qual você deseja fornecer acesso. Você pode pensar nesses grupos como grupos de perfis específicas da AWS . Qualquer usuário que seja membro desses grupos de perfis específicos recebe um único direito: acesso a um perfil específico em uma Conta da AWS específica. No entanto, esse processo único de atribuição de direitos não é escalável para gerenciar o acesso do usuário ao atribuir cada usuário a grupos de perfis específicos da AWS . Para simplificar a administração, recomendamos que você também crie vários grupos para todos os conjuntos de usuários distintos em sua organização que exigem conjuntos diferentes de AWS direitos.

Para ilustrar a configuração do IdP central, considere uma empresa com configuração do AD, em que usuários e grupos são sincronizados com o diretório do IdP. Em AWS, esses grupos do AD são mapeados para IAM funções. As principais etapas do fluxo de trabalho são as seguintes:



Fluxo de trabalho para integrar usuários, grupos e IAM funções do AD

1. Em AWS, configure a SAML integração de cada um Contas da AWS com seu IdP.
2. Em AWS, configure funções em cada uma Conta da AWS e sincronize com o IdP.
3. No sistema AD corporativo:
   a. Crie um Grupo AD para cada função da conta e sincronize com o IdP (por exemplo, `Account1-Platform-Admin-Group` (também conhecido como Grupo de AWS Funções)).
   b. Crie um grupo de gerenciamento em cada nível de personalidade (por exemplo,`Platform-Mgmt-Group`) e atribua grupos de AWS funções como membros.
   c. Atribua usuários a esse grupo de gerenciamento para permitir o acesso às Conta da AWS funções.
4. No IdP, mapeie grupos de AWS funções (como`Account1-Platform-Admin-Group`) para Conta da AWS funções (como Administrador de plataforma na Conta1).
5. Quando a cientista de dados Alice faz login no Idp, ela recebe uma interface de usuário do aplicativo AWS Federation com duas opções para escolher: “Cientista de dados da conta 1" e “cientista de dados da conta 2".
6. Alice escolhe a opção “Cientista de dados da conta 1" e eles são conectados ao aplicativo autorizado na AWS Conta 1 (console de IA). SageMaker

Para obter instruções detalhadas sobre como configurar a federação de SAML contas, consulte Como configurar SAML 2.0 para federação de AWS contas da Okta.

Federação de usuários

A autenticação para o SageMaker AI Studio pode ser feita usando IAM ou IAM iDC. Se os usuários forem gerenciados por meio deIAM, eles poderão escolher o IAM modo. Se a empresa usar um IdP externo, ela poderá se federar por meio IAM de ou IdC. IAM Observe que o modo de autenticação não pode ser atualizado para um domínio existente do SageMaker AI Studio, portanto, é fundamental tomar a decisão antes de criar um domínio do SageMaker AI Studio de produção.

Se o SageMaker AI Studio estiver configurado no IAM modo, os usuários do SageMaker AI Studio acessam o aplicativo por meio de um pré-assinado URL que conecta automaticamente o usuário ao aplicativo SageMaker AI Studio quando acessado por meio de um navegador.

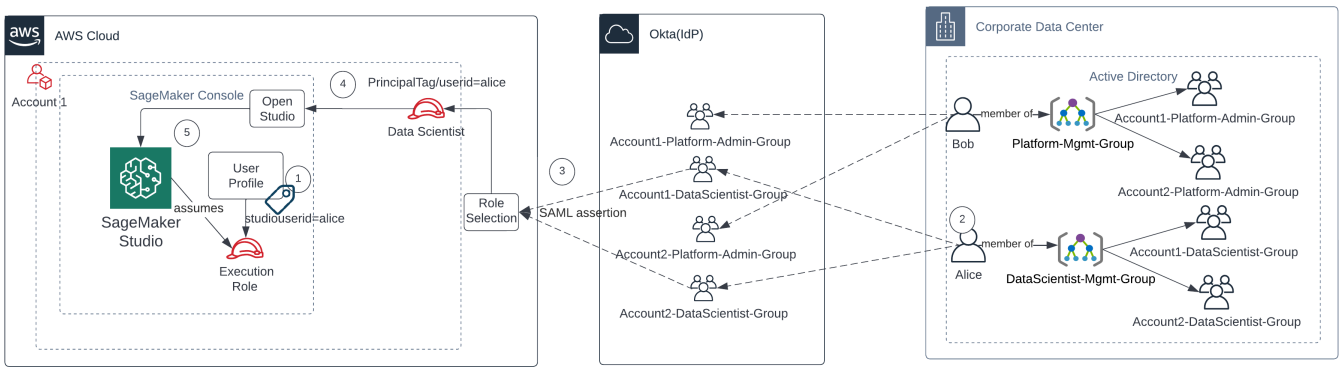
Usuários do IAM

Para IAM os usuários, o administrador cria perfis de usuário do SageMaker AI Studio para cada usuário e associa o perfil do usuário a uma IAM função que permite as ações necessárias que o usuário precisa realizar de dentro do Studio. Para impedir que um AWS usuário acesse somente seu perfil de usuário do SageMaker AI Studio, o administrador deve marcar o perfil de usuário do SageMaker AI Studio e anexar uma IAM política ao usuário que permita que ele acesse somente se o valor da tag for igual ao nome do AWS usuário. A declaração de política é assim:

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

AWS IAMou federação de contas

O método de Conta da AWS federação permite que os clientes se federem no SageMaker AI Console a partir de seu SAML IdP, como o Okta. Para impedir que os usuários acessem somente seu perfil de usuário, o administrador deve marcar o perfil de usuário do SageMaker AI Studio, adicionar `PrincipalTags` o IdP e defini-lo como tags transitivas. O diagrama a seguir mostra como o usuário federado (Data Scientist Alice) está autorizado a acessar seu próprio perfil de usuário do SageMaker AI Studio.



Acessando o SageMaker AI Studio no modo de IAM federação

1. O perfil de usuário do Alice SageMaker AI Studio é marcado com seu ID de usuário e associado à função de execução.
2. Alice se autentica no IdP (Okta).
3. O IdP autentica Alice e publica uma SAML declaração com as duas funções (Cientista de dados para as contas 1 e 2) das quais Alice é membro. Alice seleciona o perfil de cientista de dados para a conta 1.
4. Alice está conectada ao console de SageMaker IA da Conta 1, com a função assumida de Cientista de Dados. Alice abre a instância do aplicativo Studio na lista de instâncias do aplicativo Studio.
5. A tag principal de Alice na sessão de função assumida é validada em relação à tag de perfil de usuário da instância do aplicativo SageMaker AI Studio selecionada. Se a tag de perfil for válida, a instância do aplicativo SageMaker AI Studio será iniciada, assumindo a função de execução.

Se você quiser automatizar a criação de funções e políticas de execução de SageMaker IA como parte da integração do usuário, a seguir está uma maneira de fazer isso:

1. Configure um grupo do AD, como `SageMaker AI-Account1-Group` em cada conta e nível de domínio do Studio.
2. Adicione o SageMaker AI-Account1-Group à associação do grupo do usuário quando precisar integrar um usuário ao AI Studio. SageMaker

Configure um processo de automação que escute o evento de `SageMaker AI-Account1-Group` associação e use AWS APIs para criar a função, as políticas, as tags e o perfil de usuário do SageMaker AI Studio com base em suas associações ao grupo AD. Anexe o perfil ao perfil de usuário. Para obter um exemplo de política, consulteImpedir que usuários do SageMaker AI Studio acessem outros perfis de usuário.

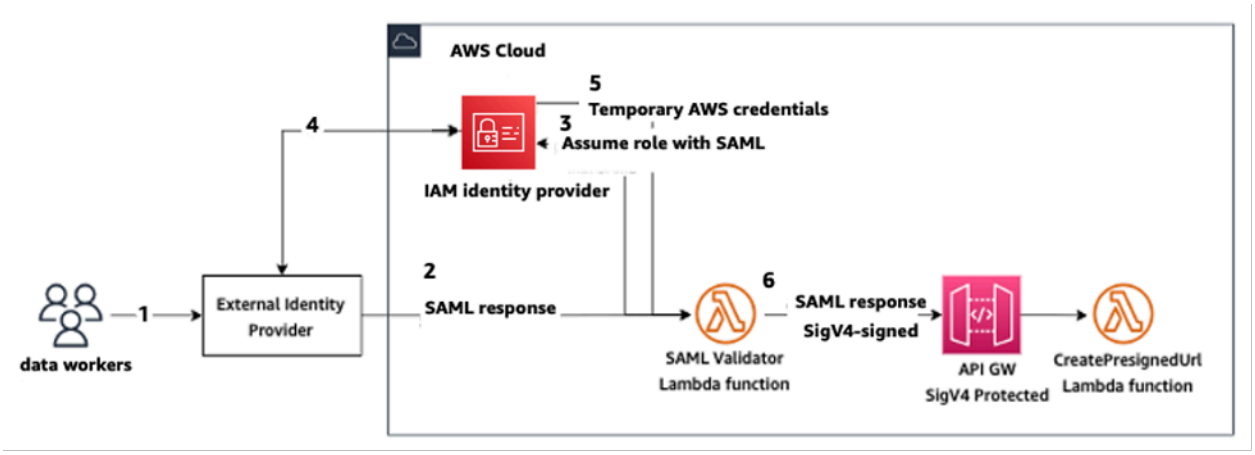
SAMLautenticação usando AWS Lambda

No IAM modo, os usuários também podem ser autenticados no SageMaker AI Studio usando SAML afirmações. Nessa arquitetura, o cliente tem um IdP existente, onde ele pode criar um SAML aplicativo para que os usuários acessem o Studio (em vez do aplicativo AWS Identity Federation). O IdP do cliente é adicionado a. IAM Uma AWS Lambda função ajuda a validar a SAML afirmação usando IAM eSTS, em seguida, invoca um gateway API ou uma função Lambda diretamente para criar o domínio pré-assinado. URL

A vantagem dessa solução é que a função Lambda pode personalizar a lógica para acesso ao SageMaker AI Studio. Por exemplo:

- Crie automaticamente um perfil de usuário se não houver um.
- Anexe ou remova funções ou documentos de política à função de execução do SageMaker AI Studio analisando os SAML atributos.
- Personalize o perfil do usuário adicionando Life Cycle Configuration (LCC) e adicionando tags.

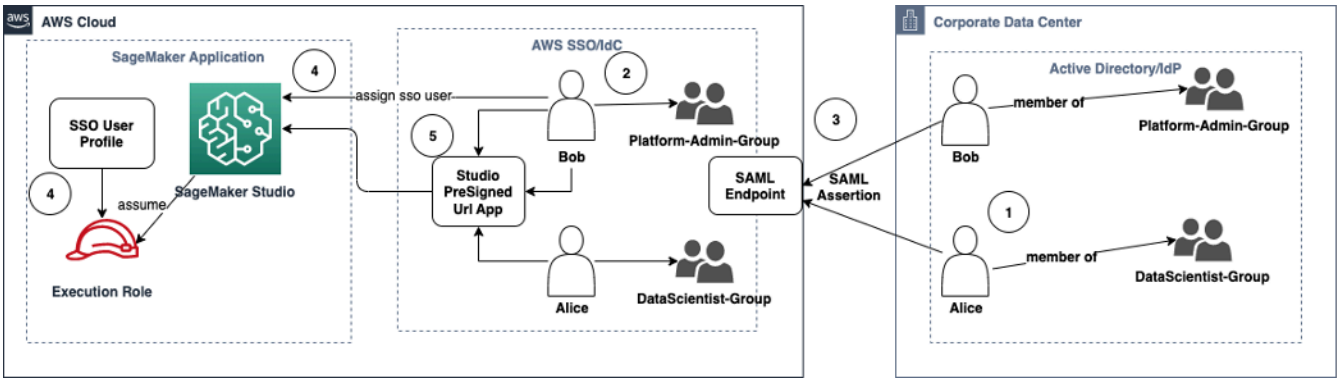
Em resumo, essa solução exporá o SageMaker AI Studio como um aplicativo SAML2 4.0 com lógica personalizada para autenticação e autorização. Consulte a seção do apêndice Acesso ao SageMaker Studio usando SAML asserção para obter detalhes de implementação.



Acessando o SageMaker AI Studio usando um SAML aplicativo personalizado

AWSIAMfederação iDC

O método de federação do IdC permite que os clientes se federem diretamente no aplicativo SageMaker AI Studio a partir de seu SAML IdP (como o Okta). O diagrama a seguir mostra como o usuário federado está autorizado a acessar sua própria instância do SageMaker AI Studio.



Acessando o SageMaker AI Studio no IAM modo iDC

1. No AD corporativo, o usuário é membro de grupos do AD, como o grupo Platform Admin e o grupo Data Scientist.
2. O usuário do AD e os grupos do AD do Provedor de Identidade (IdP) são sincronizados com o AWS IAM Identity Center e estão disponíveis como usuários e grupos de login único para atribuições, respectivamente.
3. O IdP publica uma SAML afirmação no endpoint do AWS IdC. SAML
4. No SageMaker AI Studio, o usuário do iDC é atribuído ao aplicativo SageMaker Studio. Essa tarefa pode ser feita usando o iDC Group e o SageMaker AI Studio será aplicado em cada nível de

usuário do iDC. Quando essa atribuição é criada, o SageMaker AI Studio cria o perfil de usuário do iDC e anexa a função de execução do domínio.

5. O usuário acessa o aplicativo SageMaker AI Studio usando o aplicativo seguro pré-assinado URL hospedado como um aplicativo em nuvem do iDC. SageMaker O AI Studio assume a função de execução associada ao perfil de usuário do iDC.

Orientação de autenticação de domínio

Aqui estão algumas considerações ao escolher o modo de autenticação de um domínio:

1. Se você quiser que seus usuários não acessem AWS Management Console e visualizem a interface do SageMaker AI Studio diretamente, use o modo de login único com AWS IAM o iDC.
2. Se você quiser que seus usuários não acessem AWS Management Console e visualizem a IU do SageMaker AI Studio diretamente no IAM modo, você pode fazer isso usando uma função Lambda no back-end para gerar um perfil pré-assinado URL para o usuário e redirecioná-los para a interface do AI Studio. SageMaker
3. No modo IdC, cada usuário é mapeado para um único perfil de usuário.
4. Todos os perfis de usuário recebem automaticamente o perfil de execução padrão no modo IdC. Se você quiser que seus usuários recebam funções de execução diferentes, você precisará atualizar os perfis de usuário usando UpdateUserProfileAPIo.
5. Se você quiser restringir o acesso à IU do SageMaker AI Studio no IAM modo (usando o preassinado geradoURL) a um VPC endpoint, sem atravessar a Internet, você pode usar um resolvedor personalizado. DNS Consulte a postagem pré-assinada do blog Secure Amazon SageMaker AI Studio, URLs Parte 1: Infraestrutura fundamental.

Gerenciamento de permissões

Esta seção discute as melhores práticas para configurar IAM funções, políticas e proteções comumente usadas para provisionar e operar o domínio do AI Studio. SageMaker

Funções e políticas do IAM

Como prática recomendada, talvez você queira primeiro identificar as pessoas e os aplicativos relevantes, conhecidos como diretores envolvidos no ciclo de vida do ML, e quais AWS permissões você precisa conceder a eles. Como a SageMaker IA é um serviço gerenciado, você também precisa considerar os princípios de serviço, que são AWS serviços que podem fazer API chamadas em nome de um usuário. O diagrama a seguir ilustra as diferentes IAM funções que você pode querer criar, correspondendo às diferentes personas na organização.

SageMaker IAMFunções de IA

Essas funções são descritas em detalhes, junto com alguns exemplos específicos de que IAMpermissions precisarão.

- Função de usuário do administrador de ML — Esse é o diretor que provisiona o ambiente para cientistas de dados criando domínios de estúdio e perfis de usuário (`sagemaker:CreateDomain`,`sagemaker:CreateUserProfile`), criando chaves AWS Key Management Service (AWS KMS) para usuários, criando buckets S3 para cientistas de dados e criando ECR repositórios Amazon para abrigar contêineres. Eles também podem definir configurações padrão e scripts de ciclo de vida para usuários, criar e anexar imagens personalizadas ao domínio do SageMaker AI Studio e fornecer produtos do Service Catalog, como projetos personalizados e modelos da Amazon. EMR

  Como esse diretor não executará trabalhos de treinamento, por exemplo, ele não precisa de permissões para iniciar trabalhos de treinamento ou processamento de SageMaker IA. Se eles estiverem usando a infraestrutura como modelos de código, como CloudFormation o Terraform,

para provisionar domínios e usuários, essa função seria assumida pelo serviço de provisionamento para criar os recursos em nome do administrador. Essa função pode ter acesso somente de leitura à SageMaker IA usando o. AWS Management Console

Essa função de usuário também precisará de certas EC2 permissões para iniciar o domínio em um ambiente privadoVPC, KMS permissões para criptografar o EFS volume, bem como permissões para criar uma função vinculada ao serviço para Studio (`iam:CreateServiceLinkedRole`). Descreveremos essas permissões granulares posteriormente no documento.

- Função de usuário do Data Scientist — Esse princípio é o usuário que SageMaker faz login no AI Studio, explora os dados, cria trabalhos e pipelines de processamento e treinamento, etc. A permissão principal de que o usuário precisa é a permissão para iniciar o SageMaker AI Studio, e o restante das políticas pode ser gerenciado pela função de serviço de execução de SageMaker IA.
- SageMaker Função do serviço de execução de SageMaker IA — Como a IA é um serviço gerenciado, ela lança trabalhos em nome do usuário. Esse perfil geralmente é o mais amplo em termos de permissões permitidas, porque muitos clientes optam por usar um único perfil de execução para executar trabalhos de treinamento, trabalhos de processamento ou trabalhos de hospedagem de modelos. Embora essa seja uma maneira fácil de começar, porque os clientes amadurecem em sua jornada, eles geralmente dividem a função de execução do notebook em funções separadas para API ações diferentes, especialmente ao executar essas tarefas em ambientes implantados.

  Você associa uma função ao domínio do SageMaker AI Studio na criação. No entanto, como os clientes podem precisar da flexibilidade de ter funções diferentes associadas aos diferentes perfis de usuário no domínio (por exemplo, com base em sua função profissional), você também pode associar uma IAM função separada a cada perfil de usuário. Recomendamos que você mapeie um único usuário físico para um único perfil de usuário. Se você não anexar uma função a um perfil de usuário na criação, o comportamento padrão é associar a função de execução do SageMaker AIStudio domínio também ao perfil do usuário.

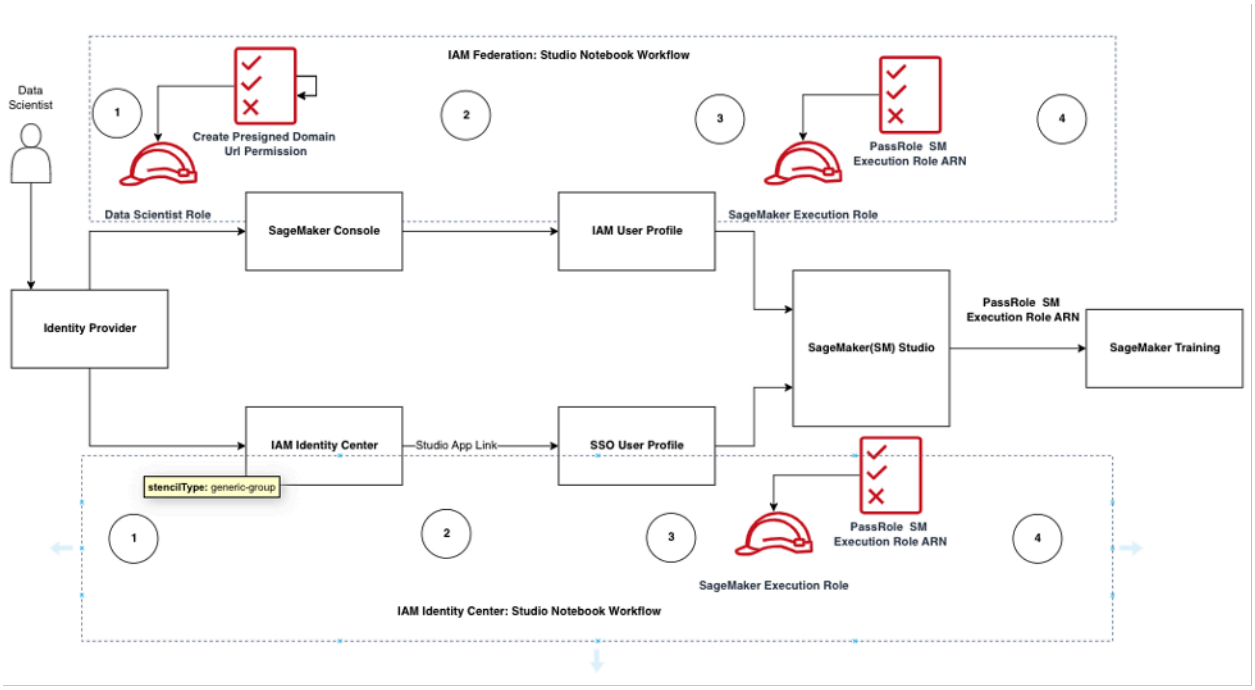
  Nos casos em que vários cientistas de dados e engenheiros de ML trabalham juntos em um projeto e precisam de um modelo de permissão compartilhado para acessar recursos, recomendamos que você crie uma função de execução de serviços de SageMaker IA em nível de equipe para compartilhar IAM as permissões entre os membros da equipe. Nos casos em que você precisa bloquear as permissões em cada nível de usuário, você pode criar uma função individual de execução de serviços de SageMaker IA em nível de usuário; no entanto, você precisa estar atento aos seus limites de serviço.

SageMaker fluxo de trabalho de autorização do AI Studio Notebook

Esta seção discute como a autorização do SageMaker AI Studio Notebook funciona para várias atividades que o cientista de dados precisa realizar para criar e treinar o modelo diretamente do SageMaker AI Studio Notebook. O domínio SageMaker AI oferece suporte a dois modos de autorização:

- IAMfederação
- IAMCentro de identidade

A seguir, este paper mostra o fluxo de trabalho de autorização do cientista de dados para cada um desses modos.



Fluxo de trabalho de autenticação e autorização para usuários do Studio

IAMFederação: fluxos de trabalho do SageMaker Studio Notebook

1. Um cientista de dados se autentica em seu provedor de identidade corporativa e assume a função de usuário de cientista de dados (a função de federação de usuários) no console de SageMaker IA. Essa função de federação tem `iam:PassRole` API permissão na função de execução de SageMaker IA para passar a função Amazon Resource Name (ARN) para o SageMaker Studio.

2. O cientista de dados seleciona o link do Open Studio em seu perfil de IAM usuário do Studio que está associado à função de execução de SageMaker IA.
3. O IDE serviço SageMaker Studio é lançado, assumindo as permissões da função de SageMaker execução do perfil do usuário. Essa função tem `iam:PassRole` API permissão para que a função de execução de SageMaker IA passe a função ARN para o serviço de treinamento de SageMaker IA.
4. Quando o Data Scientist inicia o trabalho de treinamento no (s) nó (s) de computação remota, a função de execução da SageMaker IA ARN é passada para o serviço de treinamento de SageMaker IA. Isso cria uma nova sessão de função com isso ARN e executa o trabalho de treinamento. Se precisar ampliar ainda mais a permissão para o trabalho de treinamento, você pode criar uma função específica de treinamento e passar essa função ARN ao chamar o treinamentoAPI.

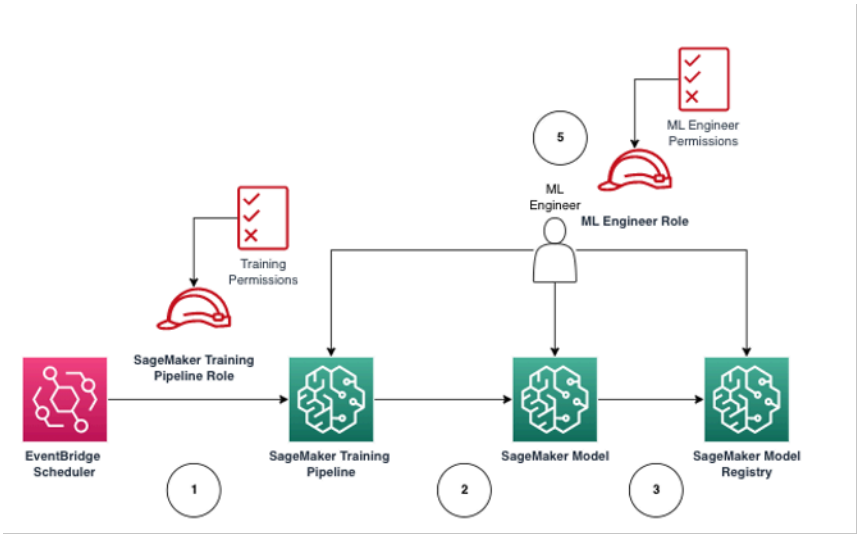
IAMIdentity Center: fluxo de trabalho do SageMaker AI Studio Notebook

1. O cientista de dados se autentica em seu provedor de identidade corporativa e clica no AWS IAM Identity Center. O cientista de dados recebe o Portal do Identity Center para o usuário.
2. O cientista de dados clica no link do aplicativo SageMaker AI Studio que foi criado a partir do perfil de usuário do iDC, que está associado à função de execução da SageMaker IA.
3. O IDE serviço SageMaker AI Studio é lançado, assumindo as permissões da função de execução de SageMaker IA do perfil do usuário. Essa função tem `iam:PassRole` API permissão para que a função de execução de SageMaker IA passe a função ARN para o serviço de treinamento de SageMaker IA.
4. Quando o cientista de dados inicia o trabalho de treinamento em nós de computação remotos, a função de execução da SageMaker IA ARN é passada para o serviço de treinamento de SageMaker IA. A função de execução ARN cria uma nova sessão de função com isso ARN e executa o trabalho de treinamento. Se precisar ampliar ainda mais a permissão para trabalhos de treinamento, você pode criar uma função específica de treinamento e aprová-la ARN ao convocar o treinamento. API

Ambiente implantado: fluxo de trabalho de treinamento de SageMaker IA

Em ambientes implantados, como testes e produção de sistemas, os trabalhos são executados por meio de agendadores automatizados e acionadores de eventos, e o acesso humano a esses

ambientes é restrito a partir dos SageMaker notebooks do AI Studio. Esta seção discute como as IAM funções funcionam com o pipeline de treinamento de SageMaker IA no ambiente implantado.



SageMaker Fluxo de trabalho de treinamento de IA em um ambiente de produção gerenciado

1. O EventBridge agendador da Amazon aciona o trabalho do pipeline de treinamento de SageMaker IA.
2. O SageMaker trabalho do pipeline de treinamento de SageMaker IA assume a função do pipeline de treinamento de IA para treinar o modelo.
3. O modelo de SageMaker IA treinado é registrado no Registro de Modelos de SageMaker IA.
4. Um engenheiro de ML assume a função de usuário engenheiro de ML para gerenciar o pipeline de treinamento e o modelo de SageMaker IA.

Permissões de dados

A capacidade dos usuários do SageMaker AI Studio de acessar qualquer fonte de dados é regida pelas permissões associadas à função de IAM execução da SageMaker IA. As políticas anexadas podem autorizá-los a ler, gravar ou excluir determinados buckets ou prefixos do Amazon S3 e a se conectar aos bancos de dados da Amazon. RDS

Acessando AWS Lake Formation dados

Muitas empresas começaram a usar data lakes governados por AWS Lake Formation para permitir o acesso refinado aos dados para seus usuários. Como exemplo desses dados controlados, os

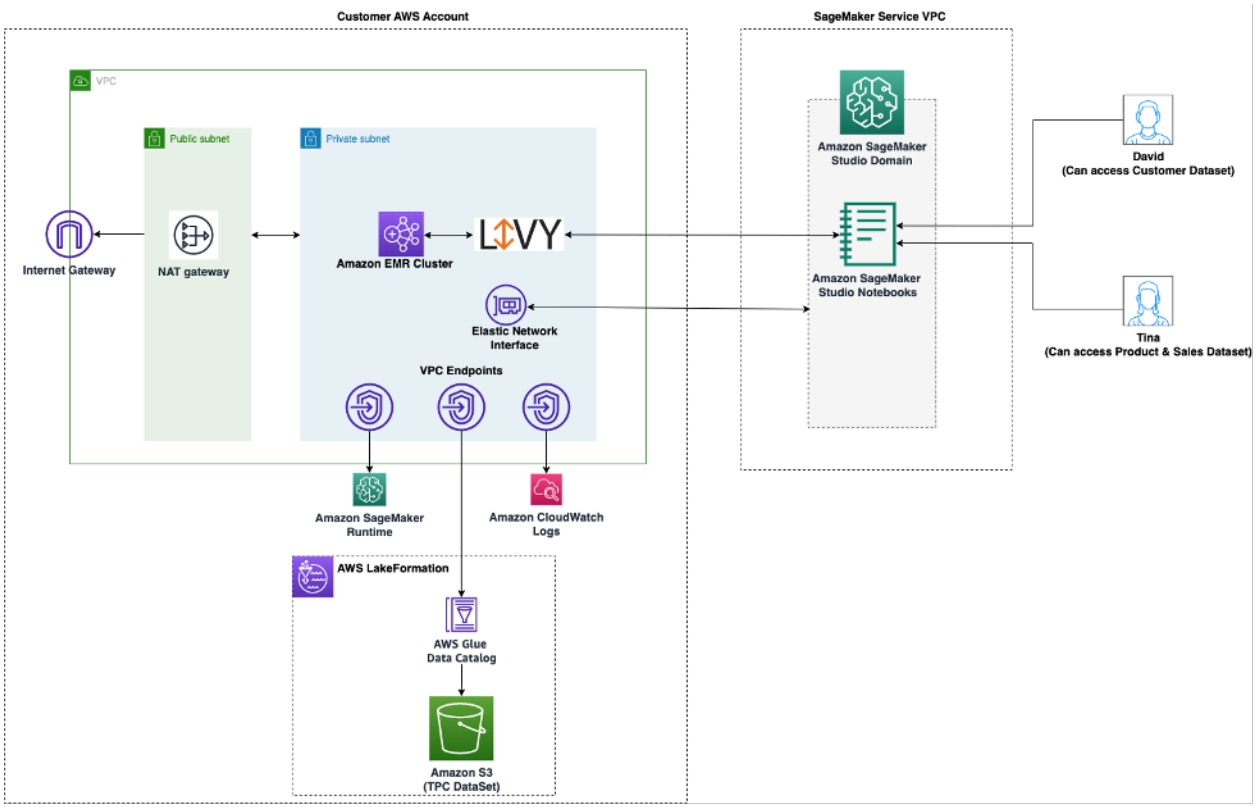
administradores podem mascarar colunas confidenciais para alguns usuários e, ao mesmo tempo, permitir consultas na mesma tabela subjacente.

Para utilizar o Lake Formation do SageMaker AI Studio, os administradores podem registrar as funções de IAM execução de SageMaker IA como`DataLakePrincipals`. Para obter mais informações, consulte a Referência de Permissões do Lake Formation. Uma vez autorizados, há três métodos principais para acessar e gravar dados controlados do SageMaker AI Studio:

1. Em um notebook SageMaker AI Studio, os usuários podem utilizar mecanismos de consulta como o Amazon Athena ou bibliotecas baseadas no boto3 para extrair dados diretamente para o notebook. A AWSSDKfor Pandas (anteriormente conhecida como awswrangler) é uma biblioteca popular. A seguir está um exemplo de código para mostrar como isso pode ser simples:

```
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. Use a conectividade nativa do SageMaker AI Studio com a Amazon EMR para ler e gravar dados em grande escala. Por meio do uso das funções de tempo de EMR execução do Apache Livy e da Amazon, o SageMaker AI Studio criou uma conectividade nativa que permite que você transfira sua IAM função de execução de SageMaker IA (ou outra função autorizada) para um EMR cluster da Amazon para acesso e processamento de dados. Consulte Connect to an Amazon EMR Cluster from Studio para up-to-date obter instruções.



Arquitetura para acessar dados gerenciados pelo Lake Formation do SageMaker Studio

3. Use a conectividade nativa do SageMaker AI Studio em sessões AWS Glue interativas para ler e gravar dados em grande escala. SageMaker Os notebooks AI Studio têm kernels integrados que permitem que os usuários executem comandos de forma interativa. AWS Glue Isso permite o uso escalável de recursos internos de Python, Spark ou Ray, que podem ler e gravar dados perfeitamente em grande escala a partir de fontes de dados controladas. Os kernels permitem que os usuários transmitam sua SageMaker execução ou outras funções autorizadasIAM. Consulte Preparar dados usando sessões AWS Glue interativas para obter mais informações.

Barreiras de proteção comuns

Esta seção discute as barreiras de proteção mais usadas para aplicar a governança em seus recursos de ML usando IAM políticas, políticas de recursos, políticas de VPC endpoint e políticas de controle de serviços (). SCPs

Limitar o acesso ao notebook a instâncias específicas

Essa política de controle de serviços pode ser usada para limitar os tipos de instância aos quais os cientistas de dados têm acesso ao criar notebooks do Studio. Observe que qualquer usuário

precisará da instância de “sistema” permitida para criar o aplicativo Jupyter Server padrão que hospeda o SageMaker AI Studio.

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

Limite os domínios não compatíveis do SageMaker AI Studio

Para domínios do SageMaker AI Studio, a política de controle de serviço a seguir pode ser usada para forçar o tráfego a acessar os recursos do cliente, de forma que eles não passem pela Internet pública, mas pela do cliente: VPC

```
{
     "Version": "2012-10-17",
     "Statement": [
         {   "Sid": "LockDownStudioDomain",
             "Effect": "Deny",
             "Action": [
                 "sagemaker:CreateDomain"
             ],
             "Resource": "*",
```

```
            "Condition": {
                        "StringNotEquals": {"sagemaker:AppNetworkAccessType":
"VpcOnly"
                },
                "Null": {
                        "sagemaker:VpcSubnets": "true",
                        "sagemaker:VpcSecurityGroupIds": "true"
                }
            }
        }
    ]
}
```

Limite o lançamento de imagens de SageMaker IA não autorizadas

A política a seguir impede que um usuário inicie uma imagem de SageMaker IA não autorizada em seu domínio:f

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

Inicie notebooks somente por meio de endpoints de SageMaker IA VPC

Além dos VPC endpoints para o plano de controle de SageMaker IA, a SageMaker IA oferece suporte a VPC endpoints para que os usuários se conectem aos notebooks do SageMaker AI Studio ou às instâncias do notebook SageMaker AI. Se você já configurou um VPC endpoint para uma instância do SageMaker AI Studio/Notebook, a chave de IAM condição a seguir só permitirá conexões com notebooks do SageMaker AI Studio se elas forem feitas por meio do endpoint do SageMaker AI Studio ou do VPC endpoint do AI. SageMaker API

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

Limite o acesso ao notebook SageMaker AI Studio a uma faixa limitada de IP

As empresas geralmente limitam o acesso ao SageMaker AI Studio a determinados intervalos de IP corporativos permitidos. A IAM política a seguir com a chave de `SourceIP` condição pode limitar isso.

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

Impedir que usuários do SageMaker AI Studio acessem outros perfis de usuário

Como administrador, ao criar o perfil de usuário, certifique-se de que o perfil esteja marcado com o nome de usuário do SageMaker AI Studio com a chave de tag`studiouserid`. A entidade principal (usuário ou perfil associada ao usuário) também deve ter uma tag com a chave `studiouserid` (essa tag pode ter qualquer nome e não está restrita a `studiouserid`).

Em seguida, anexe a política a seguir à função que o usuário assumirá ao iniciar o SageMaker AI Studio.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "AmazonSageMakerPresignedUrlPolicy",
            "Effect": "Allow",
            "Action": [
                "sagemaker:CreatePresignedDomainUrl"
            ],
            "Resource": "*",
```

```
            "Condition": {
                "StringEquals": {
                    "sagemaker:ResourceTag/studiouserid": "${aws:PrincipalTag/
studiouserid}"
                }
            }
        }
    ]
}
```

Garantir a marcação

Os cientistas de dados precisam usar os notebooks do SageMaker AI Studio para explorar dados, criar e treinar modelos. A aplicação de etiquetas em notebooks ajuda a monitorar o uso e controlar os custos, além de garantir a propriedade e a auditabilidade.

Para aplicativos do SageMaker AI Studio, verifique se o perfil do usuário está marcado. As tags são propagadas automaticamente para os aplicativos a partir do perfil do usuário. Para impor a criação de perfil de usuário com tags (suportadas por CLI eSDK), considere adicionar essa política à função de administrador:

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

Para outros recursos, como trabalhos de treinamento e trabalhos de processamento, você pode tornar as tags obrigatórias usando a seguinte política:

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

Acesso root no SageMaker AI Studio

No SageMaker AI Studio, o notebook é executado em um contêiner Docker que, por padrão, não tem acesso root à instância hospedeira. Da mesma forma, além do usuário run-as padrão, todos os outros intervalos de IDs de usuário dentro do contêiner são remapeados como usuário não privilegiado na própria instância host. IDs Como resultado, a ameaça de escalonamento de privilégios é limitada ao próprio contêiner do notebook.

Ao criar imagens personalizadas, talvez você queira fornecer ao usuário permissões não root para controles mais rígidos; por exemplo, evitar executar processos indesejáveis como root ou instalar pacotes disponíveis publicamente. Nesses casos, você pode criar a imagem para ser executada como usuário não raiz no Dockerfile. Se você criar o usuário como root ou não root, você precisa garantir que o UID/GID of the user is identical to the UID/GID seja AppImageConfigpara o aplicativo personalizado, que cria a configuração para que a SageMaker IA execute um aplicativo usando a imagem personalizada. Por exemplo, se seu Dockerfile for criado para um usuário não raiz, como o seguinte:

```
ARG NB_UID="1000"
```

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

O `AppImageConfig` arquivo precisa mencionar o mesmo UID e GID em seu`KernelGatewayConfig`:

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

Os GID valoresUID/aceitáveis para imagens personalizadas são 0/0 e 1000/100 para imagens do Studio. Para exemplos de criação de imagens personalizadas e as configurações associadas `AppImageConfig`, consulte este repositório do Github.

Para evitar que os usuários adulterem isso, não conceda as `DeleteAppImageConfig` permissões `CreateAppImageConfigUpdateAppImageConfig`, ou aos usuários do notebook SageMaker AI Studio.

Gerenciamento de rede

Para configurar o domínio do SageMaker AI Studio, você precisa especificar a VPC rede, as sub-redes e os grupos de segurança. Ao especificar as sub-redes VPC e, assegure-se de alocar IPs considerando o volume de uso e o crescimento esperado, discutidos nas seções a seguir.

VPCplanejamento de rede

As VPC sub-redes do cliente associadas ao domínio do SageMaker AI Studio devem ser criadas com o intervalo apropriado de roteamento entre domínios sem classe (CIDR), dependendo dos seguintes fatores:

- Número de usuários.
- Número de aplicativos por usuário.
- Número de tipos de instância exclusivos por usuário.
- Número médio de instâncias de treinamento por usuário.
- Porcentagem de crescimento esperada.

SageMaker A IA e AWS os serviços participantes injetam interfaces de rede elásticas (ENI) na VPC sub-rede do cliente para os seguintes casos de uso:

- A Amazon EFS injeta um destino ENI for an EFS mount para o domínio de SageMaker IA (um IP por sub-rede/zona de disponibilidade anexada ao SageMaker domínio de IA).
- SageMaker O AI Studio injeta um ENI para cada instância exclusiva usada por um perfil de usuário ou por um espaço compartilhado. Por exemplo:
  - Se um perfil de usuário executa um aplicativo de servidor Jupyter padrão (uma instância de “sistema”), um aplicativo Data Science e um aplicativo Base Python (ambos executados em uma instância `ml.t3.medium`), o Studio injeta dois endereços IP.
  - Se um perfil de usuário executa um aplicativo de servidor Jupyter padrão (uma instância de “sistema”), um GPU aplicativo Tensorflow (em uma `ml.g4dn.xlarge` instância) e um aplicativo de processamento de dados (em uma `ml.m5.4xlarge` instância), o Studio injeta três endereços IP.
- Um ENI para cada VPC endpoint nas VPC sub-redes/zonas de disponibilidade do domínio é injetado (quatro IPs para endpoints de SageMaker IA; aproximadamente seis IPs para VPC endpoints de serviços participantes, como S3, VPC e.) ECR CloudWatch

- Se os trabalhos de treinamento e processamento de SageMaker IA forem iniciados com a mesma VPC configuração, cada trabalho precisará de dois endereços IP por instância.

> **Note**
>
> VPCas configurações do SageMaker AI Studio, como sub-redes e tráfego VPC somente, não são repassadas automaticamente para as tarefas de treinamento/processamento criadas a partir do AI Studio. SageMaker O usuário precisa definir VPC as configurações e o isolamento da rede conforme necessário ao chamar o APIs Create*Job. Consulte Executar contêineres de treinamento e inferência executados no modo sem Internet para maiores informações.

Cenário: um cientista de dados realiza experimentos em dois tipos de instância diferentes

Nesse cenário, suponha que um domínio de SageMaker IA esteja configurado VPC somente no modo de tráfego. Existem VPC endpoints configurados, como SageMaker AIAPI, SageMaker AI runtime, Amazon S3 e Amazon. ECR

Um cientista de dados está realizando experimentos em notebooks Studio, executando em dois tipos de instância diferentes (por exemplo, `ml.t3.medium` e `ml.m5.large`) e lançando dois aplicativos em cada tipo de instância.

Suponha que o cientista de dados também esteja executando simultaneamente um trabalho de treinamento com a mesma VPC configuração em uma `ml.m5.4xlarge` instância.

Nesse cenário, o serviço SageMaker AI Studio injetará da ENIs seguinte forma:

Tabela 1 — ENIs injetada no cliente VPC para um cenário de experimentação

Entidade	Alvo	ENIinjetado	Observações	Nível
EFSalvo de montagem	VPCsub-redes	Três	Três AZs /sub-redes	Domínio
Endpoints do VPC	VPCsub-redes	30	Três AZs /sub-redes com 10 cada VPCE	Domínio

Entidade	Alvo	ENIinjetado	Observações	Nível
Servidor Jupyter	Sub-rede VPC	Um	Um IP por instância	Usuário
KernelGateway Aplicativo	Sub-rede VPC	Dois	Um IP por tipo de instância	Usuário
Treinamento	Sub-rede VPC	Dois	Dois IPs por instância de treinamento Cinco IPs por instância de treinamento, se EFAfor usado	Usuário

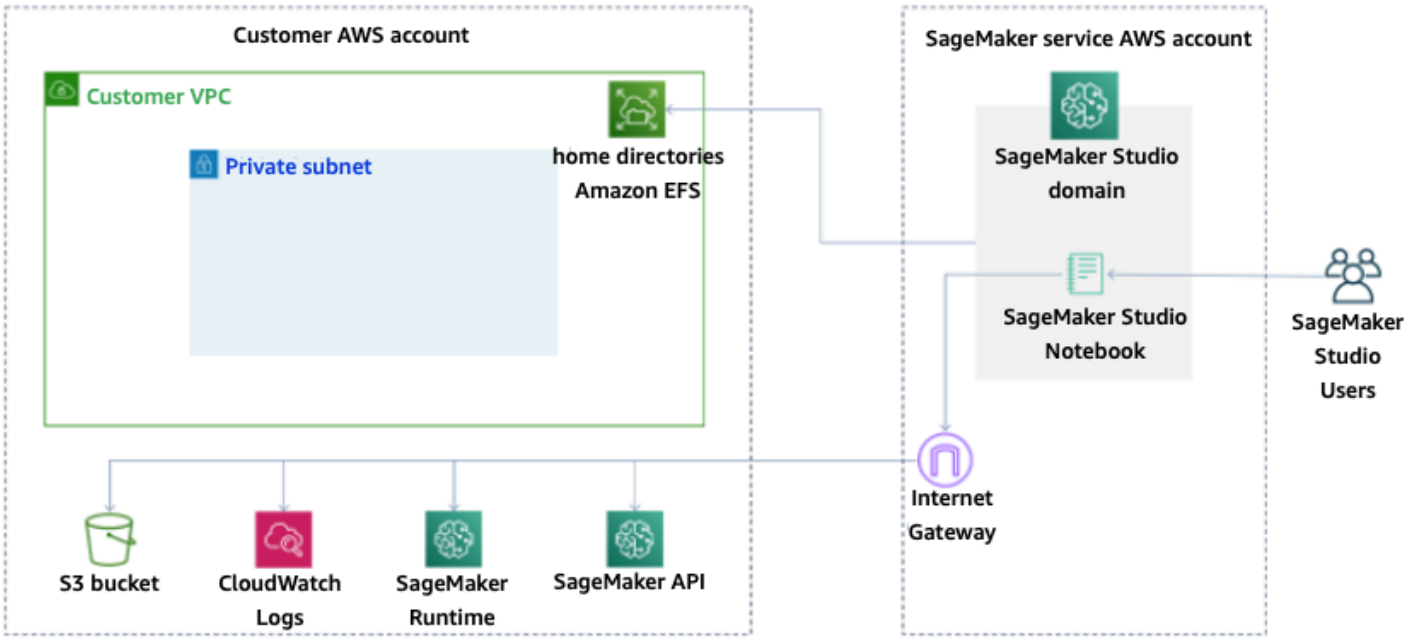
Nesse cenário, há um total de 38 IPs consumidos no cliente, VPC IPs sendo 33 compartilhados entre usuários no nível do domínio e cinco IPs são consumidos no nível do usuário. Se você tiver 100 usuários com perfis de usuário semelhantes nesse domínio realizando essas atividades simultaneamente, você consumirá cinco x 100 = 500 IPs no nível do usuário, além do consumo de IP no nível do domínio, que é 11 IPs por sub-rede, totalizando 511. IPs Para esse cenário, você precisa criar a VPC sub-rede CIDR com /22 que alocará 1024 endereços IP, com espaço para crescer.

VPCopções de rede

Um domínio do SageMaker AI Studio oferece suporte à configuração da VPC rede com uma das seguintes opções:

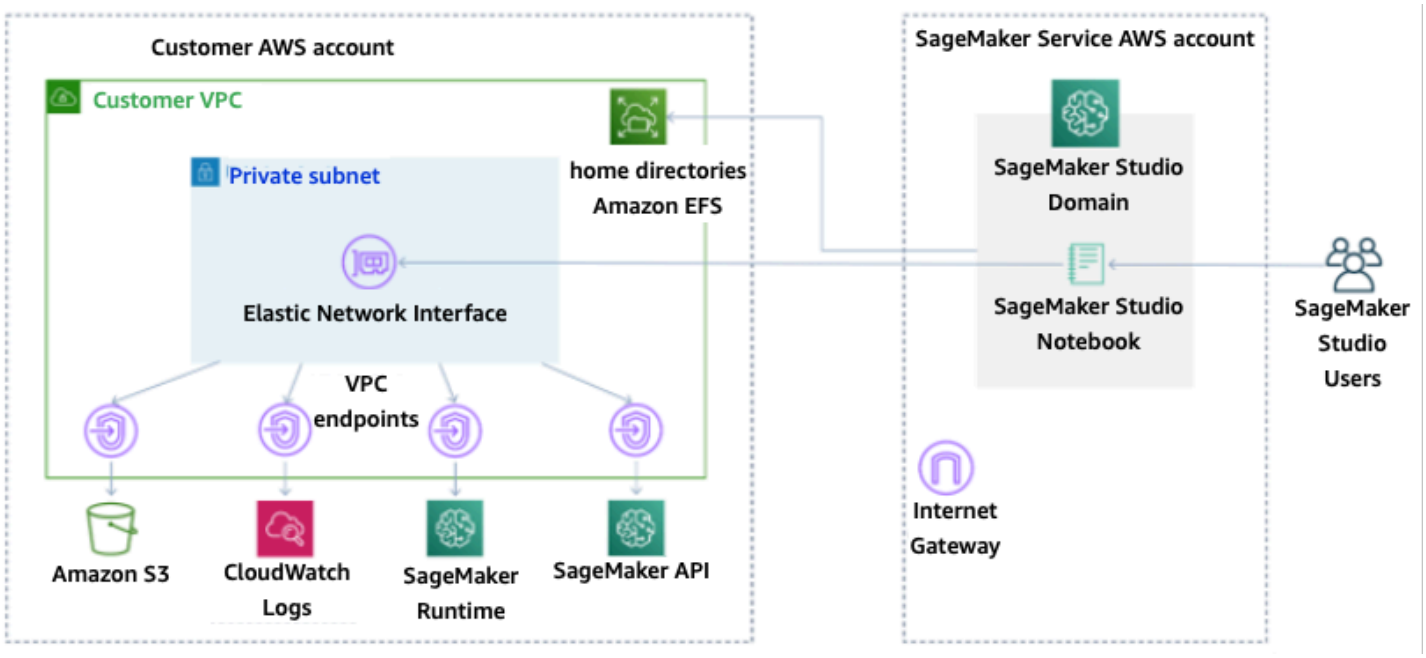
- Somente internet pública
- Somente VPC

A opção somente de Internet pública permite que API os serviços de SageMaker IA usem a Internet pública por meio do gateway de Internet provisionado noVPC, gerenciado pela conta de serviço de SageMaker IA, conforme mostrado no diagrama a seguir:



Modo padrão: acesso à Internet via conta de serviço SageMaker AI

A VPCúnica opção desativa o roteamento da Internet a partir da conta de serviço VPC gerenciada pela SageMaker IA e permite que o cliente configure o tráfego a ser roteado pelos VPC endpoints, conforme mostrado no diagrama a seguir:



VPCúnico modo: sem acesso à Internet por meio da conta de serviço de SageMaker IA

Para um domínio configurado no modo VPC único, configure um grupo de segurança por perfil de usuário para garantir o isolamento completo das instâncias subjacentes. Cada domínio em uma AWS conta pode ter sua própria VPC configuração e modo de internet. Para obter mais detalhes sobre a configuração da VPC rede, consulte Connect SageMaker AI Studio Notebooks in a VPC to External Resources.

Limitações

- Depois que um domínio do SageMaker AI Studio é criado, você não pode associar novas sub-redes ao domínio.
- O tipo de VPC rede (somente internet pública ou VPCsomente) não pode ser alterado.

Proteção de dados

Antes de arquitetar uma carga de trabalho de ML, as práticas básicas que influenciam a segurança devem estar em vigor. Por exemplo, a classificação de dados fornece uma forma de categorizar os dados com base nos níveis de sensibilidade, e a criptografia protege os dados, tornando-os ininteligíveis para acesso não autorizado. Esses métodos são importantes porque apoiam objetivos como evitar o manuseio indevido ou o cumprimento de obrigações regulatórias.

SageMaker O AI Studio fornece vários recursos para proteger dados em repouso e em trânsito. No entanto, conforme descrito no modelo de Responsabilidade AWS Compartilhada, os clientes são responsáveis por manter o controle sobre o conteúdo hospedado na infraestrutura AWS global. Nesta seção, descreveremos como os clientes podem usar esses recursos para proteger seus dados.

Proteja dados em repouso

Para proteger seus notebooks do SageMaker AI Studio junto com seus dados de construção de modelos e artefatos de modelo, a SageMaker IA criptografa os notebooks, bem como a saída dos trabalhos de treinamento e transformação em lote. SageMaker A IA os criptografa por padrão, usando a chave AWS gerenciada para o Amazon S3. Essa chave AWS gerenciada para o Amazon S3 não pode ser compartilhada para acesso entre contas. Para acesso entre contas, especifique sua chave gerenciada pelo cliente ao criar recursos de SageMaker IA para que ela possa ser compartilhada para acesso entre contas.

Com o SageMaker AI Studio, os dados podem ser armazenados nos seguintes locais:

- Bucket S3 — Quando um notebook compartilhável é ativado, o SageMaker AI Studio compartilha instantâneos e metadados do notebook em um bucket S3.
- EFSvolume — O SageMaker AI Studio anexa um EFS volume ao seu domínio para armazenar cadernos e arquivos de dados. Esse EFS volume persiste mesmo depois que o domínio é excluído.
- EBSvolume — EBS é anexado à instância em que o notebook é executado. Esse volume persiste durante a instância.

Criptografia em repouso com AWS KMS

- Você pode passar sua AWS KMS chave para criptografar um EBS volume anexado a notebooks, treinamentos, ajustes, trabalhos de transformação em lote e endpoints.
- Se você não especificar uma KMS chave, a SageMaker IA criptografa os volumes do sistema operacional (OS) e os volumes de dados de ML com uma chave gerenciada pelo sistemaKMS.
- Dados confidenciais que precisam ser criptografados com uma KMS chave por motivos de conformidade devem ser armazenados no volume de armazenamento de ML ou no Amazon S3. Ambos podem ser criptografados usando uma KMS chave especificada por você.

Proteger dados em trânsito

SageMaker O AI Studio garante que os artefatos do modelo de ML e outros artefatos do sistema sejam criptografados em trânsito e em repouso. As solicitações para a SageMaker IA API e o console são feitas por meio de uma conexão segura (SSL). Alguns dados dentro da rede em trânsito (dentro da plataforma de serviço) não são criptografados. Isso inclui:

- Comunicações de comando e controle entre o plano de controle de serviço e as instâncias de trabalho de treinamento (não dados do cliente).
- Comunicações entre nós em trabalhos de treinamento processamento distribuídos (dentro da rede).

No entanto, você pode optar por criptografar a comunicação entre os nós em um cluster de treinamento. A habilitação da criptografia de tráfego entre contêineres pode aumentar o tempo de treinamento, especialmente se você estiver usando algoritmos de aprendizado profundo distribuídos.

Por padrão, a Amazon SageMaker AI executa trabalhos de treinamento em uma Amazon VPC para ajudar a manter seus dados seguros. Você pode adicionar outro nível de segurança para proteger seus contêineres e dados de treinamento configurando um ambiente privadoVPC. Além disso, você pode configurar seu domínio do SageMaker AI Studio para ser executado VPC somente no modo e configurar VPC endpoints para rotear o tráfego por uma rede privada sem gerar tráfego pela Internet.

Barreiras de proteção de dados

Criptografe volumes de hospedagem de SageMaker IA em repouso

Use a política a seguir para aplicar a criptografia durante a hospedagem de um endpoint de SageMaker IA para inferência on-line:

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

Criptografe buckets S3 usados durante o monitoramento de modelos

O Model Monitoring captura dados enviados para seu endpoint de SageMaker IA e os armazena em um bucket S3. Ao configurar o Data Capture Config, você precisa criptografar o bucket do S3. Atualmente, não há controle compensatório para isso.

Além de capturar as saídas do endpoint, o serviço Model Monitoring verifica o desvio em relação a uma linha de base pré-especificada. Você precisa criptografar as saídas e os volumes intermediários de armazenamento usados para monitorar o desvio.

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

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

Criptografar um volume de armazenamento de domínio do SageMaker AI Studio

Aplique criptografia ao volume de armazenamento anexado ao domínio do Studio. Essa política exige que o usuário forneça um CMK para criptografar os volumes de armazenamento anexados aos domínios do estúdio.

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

Criptografe dados armazenados no S3 que são usados para compartilhar notebooks

Essa é a política para criptografar todos os dados armazenados no bucket que são usados para compartilhar notebooks entre usuários em um domínio do SageMaker AI Studio:

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

Limitações

- Depois que um domínio é criado, você não pode atualizar o armazenamento de EFS volume anexado com uma AWS KMS chave personalizada.
- Você não pode atualizar tarefas de treinamento/processamento ou configurações de endpoints com KMS chaves depois de criadas.

Registro em log e monitoramento

Para ajudá-lo a depurar seus trabalhos de compilação, trabalhos de processamento, trabalhos de treinamento, endpoints, trabalhos de transformação, instâncias de notebook e configurações de ciclo de vida de instâncias de notebook, tudo o que um contêiner de algoritmo, um contêiner de modelo ou uma configuração de ciclo de vida de instância de notebook envia para stdout ou stderr também é enviado para o Amazon Logs. CloudWatch Você pode monitorar o SageMaker AI Studio usando a Amazon CloudWatch, que coleta dados brutos e os processa em métricas legíveis, quase em tempo real. Essas estatísticas são mantidas por 15 meses, de maneira que você possa acessar informações históricas e ter uma perspectiva melhor de como o aplicativo web ou o serviço está se saindo.

Fazendo login com CloudWatch

Como o processo de ciência de dados é inerentemente experimental e iterativo, é essencial registrar atividades como uso do notebook, tempo de execução do trabalho de treinamento/processamento, métricas de treinamento e métricas de atendimento de endpoints, como latência de invocação. Por padrão, a SageMaker IA publica métricas no CloudWatch Logs, e esses registros podem ser criptografados com chaves gerenciadas pelo cliente usando. AWS KMS

Você também pode usar VPC endpoints para enviar registros CloudWatch sem usar a Internet pública. Você também pode definir alarmes que observam determinados limites e enviam notificações ou realizam ações quando esses limites são atingidos. Para obter mais informações, consulte o Guia do CloudWatch usuário da Amazon.

SageMaker A IA cria um único grupo de registros para o Studio, em`/aws/sagemaker/studio`. Cada perfil de usuário e aplicativo tem seu próprio fluxo de logs nesse grupo de logs, e os scripts de configuração do ciclo de vida também têm seu próprio fluxo de logs. Por exemplo, um perfil de usuário chamado “studio-user” com um aplicativo Jupyter Server e com um script de ciclo de vida anexado, e um aplicativo Data Science Kernel Gateway tem os seguintes fluxos de logs:

```
/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
```

Para que a SageMaker IA envie registros CloudWatch em seu nome, o chamador do Training/Processing/Transform trabalho APIs precisará das seguintes permissões:

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

Para criptografar esses registros com uma AWS KMS chave personalizada, primeiro você precisará modificar a política de chaves para permitir que o CloudWatch serviço criptografe e descriptografe a chave. Depois de criar uma AWS KMS chave de criptografia de log, modifique a política de chaves para incluir o seguinte:

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

Observe que você sempre pode usar `ArnEquals` e fornecer um Amazon Resource Name (ARN) específico para o CloudWatch log que você deseja criptografar. Aqui, mostramos que você pode usar essa chave para criptografar todos os logs em uma conta para simplificar. Além disso, endpoints de treinamento, processamento e modelo publicam métricas sobre a utilização da instância CPU e da memória, a latência da invocação de hospedagem e assim por diante. Você também pode configurar SNS a Amazon para notificar os administradores sobre eventos quando determinados limites forem ultrapassados. O consumidor do treinamento e do processamento APIs precisa ter as seguintes permissões:

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

Auditoria com AWS CloudTrail

Para melhorar sua postura de conformidade, audite tudo o que você APIs precisa. AWS CloudTrail Por padrão, todas as SageMaker IA APIs são registradas com AWS CloudTrail. Você não precisa de nenhuma IAM permissão adicional para habilitar CloudTrail.

Todas as ações de SageMaker IA, com exceção de `InvokeEndpoint` e`InvokeEndpointAsync`, são registradas CloudTrail e documentadas nas operações. Por exemplo, chamadas para as `CreateNotebookInstance` ações `CreateTrainingJob`e `CreateEndpoint`, e geram entradas nos arquivos de CloudTrail log.

Cada entrada de CloudTrail evento contém informações sobre quem gerou a solicitação. As informações de identidade ajudam a determinar o seguinte:

- Se a solicitação foi feita com credenciais de usuário da raiz ou do AWS IAM.
- Se a solicitação foi feita com credenciais de segurança temporárias de uma função ou de um usuário federado.
- Se a solicitação foi feita por outro AWS serviço. Para ver um exemplo de evento, consulte o Log SageMaker AI API Calls com a CloudTrail documentação.

Por padrão, CloudTrail registra o nome da função de execução do Studio do perfil do usuário como identificador de cada evento. Isso funciona se cada usuário tiver seu próprio perfil de execução. Se vários usuários compartilharem a mesma função de execução, você poderá usar a `sourceIdentity` configuração para propagar o nome do perfil de usuário do Studio para CloudTrail. Consulte Monitoramento do acesso aos recursos do usuário do Amazon SageMaker AI Studio para ativar o `sourceIdentity` recurso. Em um espaço compartilhado, todas as ações se referem ao espaço ARN como fonte, e você não pode auditá-las`sourceIdentity`.

Atribuição de custos

SageMaker O AI Studio tem recursos integrados para ajudar os administradores a monitorar os gastos de seus domínios individuais, espaços compartilhados e usuários.

Marcação automática

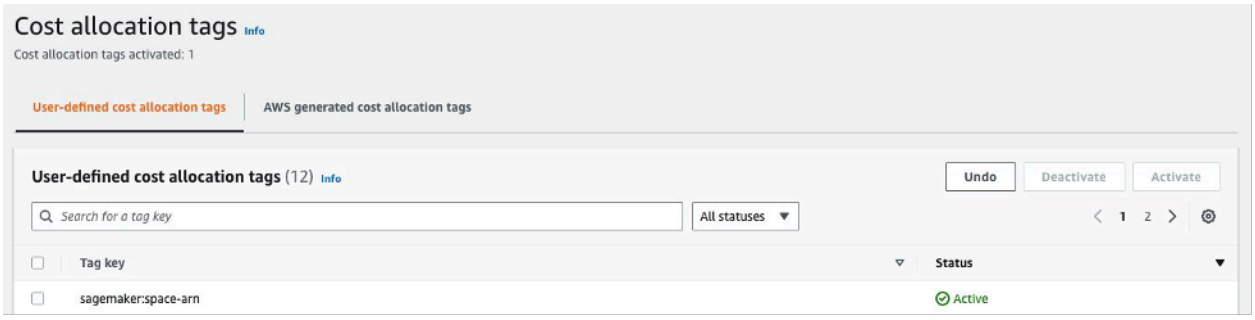
SageMaker Agora, o AI Studio marca automaticamente novos SageMaker recursos, como trabalhos de treinamento, trabalhos de processamento e aplicativos do kernel, com seus respectivos`sagemaker:domain-arn`. Em um nível mais granular, a SageMaker IA também marca o recurso com o `sagemaker:user-profile-arn` ou `sagemaker:space-arn` para designar o principal criador do recurso.

SageMaker Os EFS volumes de domínio de IA são marcados com uma chave nomeada `ManagedByAmazonSageMakerResource` com o valor do domínioARN. Eles não têm tags granulares para entender o uso do espaço em um nível por usuário. No entanto, os administradores podem conectar o EFS volume a uma EC2 instância para monitoramento personalizado.

Monitoramento de custos

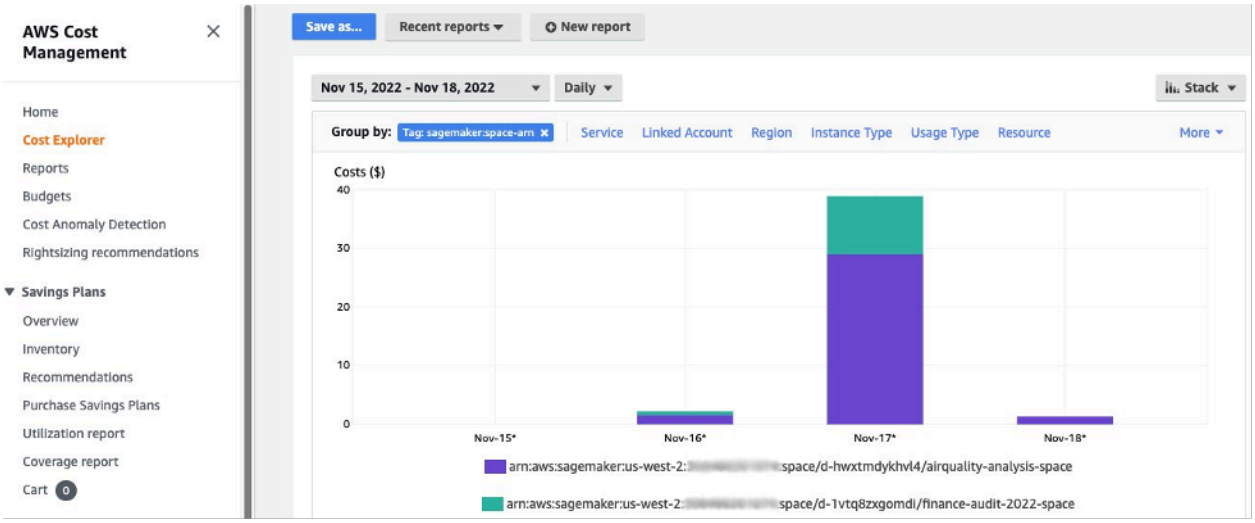
As tags automatizadas permitem que os administradores rastreiem, relatem e monitorem seus gastos com ML por meio de out-of-the-box soluções como AWS Cost Explorere AWS Budgets, bem como soluções personalizadas criadas com base nos dados dos Relatórios de AWS Custo e Uso (CURs).

Para usar as tags anexadas para análise de custos, elas devem primeiro ser ativadas na seção Tags de alocação de custos do console AWS Billing . Pode levar até 24 horas para que as tags apareçam no painel de tags de alocação de custos, então você precisará criar um recurso de SageMaker IA antes de ativá-las.



Espaço ARN ativado como tags de alocação de custos no Cost Explorer

Depois de ativar uma tag de alocação de custos, AWS começará a rastrear seus recursos marcados e, após 24 a 48 horas, as tags aparecerão como filtros selecionáveis no Cost Explorer.



Custos agrupados por espaço compartilhado para um domínio de amostra

Controle de custos

Quando o primeiro usuário do SageMaker AI Studio é integrado, a SageMaker IA cria um EFS volume para o domínio. Os custos de armazenamento desse EFS volume são incorridos, pois notebooks e arquivos de dados são armazenados no diretório inicial do usuário. Quando o usuário inicia os notebooks Studio, eles são executados para as instâncias computacionais que executam os notebooks. Consulte os preços do Amazon SageMaker AI para obter uma análise detalhada dos custos.

Os administradores podem controlar os custos de computação especificando a lista de instâncias que um usuário pode criar, usando IAM as políticas mencionadas na seção Guardrails comuns. Além disso, recomendamos que os clientes usem a extensão de desligamento automático do SageMaker AI Studio para economizar custos ao desligar automaticamente os aplicativos inativos. Essa extensão de servidor pesquisa periodicamente os aplicativos em execução por perfil de usuário e desliga os aplicativos ociosos com base em um tempo limite definido pelo administrador.

Para definir essa extensão para todos os usuários em seu domínio, você pode usar uma configuração de ciclo de vida conforme descrito na seção Personalização. Além disso, você também pode usar o verificador de extensão para garantir que todos os usuários do seu domínio tenham a extensão instalada.

Personalização

Configuração do ciclo de vida

As configurações de ciclo de vida são scripts de shell iniciados por eventos de ciclo de vida do SageMaker AI Studio, como iniciar um novo notebook do AI Studio. SageMaker Você pode usar esses scripts de shell para automatizar a personalização de seus ambientes do SageMaker AI Studio, como instalar pacotes personalizados, a extensão Jupyter para desligamento automático de aplicativos de notebook inativos e definir a configuração do Git. Para obter instruções detalhadas sobre como criar configurações de ciclo de vida, consulte este blog: Personalize o Amazon SageMaker AI Studio usando configurações de ciclo de vida.

Imagens personalizadas para notebooks SageMaker AI Studio

Os notebooks Studio vêm com um conjunto de imagens pré-criadas, que consistem no Amazon AI SageMaker Python SDK e na versão mais recente do runtime ou kernel. IPython Com esse recurso, você pode trazer suas próprias imagens personalizadas para os notebooks Amazon SageMaker AI. Essas imagens ficam então disponíveis para todos os usuários autenticados no domínio.

Desenvolvedores e cientistas de dados podem precisar de imagens personalizadas para vários casos de uso diferentes:

- Acesso a versões específicas ou mais recentes de estruturas de ML populares TensorFlow, como, MXNet PyTorch, ou outras.
- Leve códigos ou algoritmos personalizados desenvolvidos localmente aos notebooks do SageMaker AI Studio para acelerar a iteração e o treinamento de modelos.
- Acesso a lagos de dados ou armazenamentos de dados locais viaAPIs. Os administradores precisam incluir os drivers correspondentes na imagem.
- Acesso a um tempo de execução de back-end (também chamado de kernel), diferente de IPython (como R, Julia ou outros). Também é possível usar a abordagem descrita para instalar um kernel personalizado.

Para obter instruções detalhadas sobre como criar uma imagem personalizada, consulte Criar uma imagem de SageMaker IA personalizada.

JupyterLab extensões

Com o SageMaker AI Studio JuypterLab 3 Notebook, você pode aproveitar a crescente comunidade de extensões de código aberto JupyterLab. Esta seção destaca algumas que se encaixam naturalmente no fluxo de trabalho do desenvolvedor de SageMaker IA, mas recomendamos que você procure as extensões disponíveis ou até mesmo crie suas próprias.

JupyterLab O 3 agora facilita significativamente o processo de empacotamento e instalação de extensões. Você pode instalar as extensões mencionadas acima por meio de scripts bash. Por exemplo, no SageMaker AI Studio, abra o terminal do sistema a partir do inicializador do Studio e execute os seguintes comandos. Além disso, você pode automatizar a instalação dessas extensões usando configurações de ciclo de vida para que elas persistam entre as reinicializações do Studio. Você pode configurar isso para todos os usuários no domínio ou em um nível de usuário individual.

Por exemplo, para instalar uma extensão para um navegador de arquivos Amazon S3, execute os seguintes comandos no terminal do sistema e certifique-se de atualizar seu navegador:

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

Para obter mais informações sobre gerenciamento de extensões, incluindo como criar configurações de ciclo de vida que funcionem para as versões 1 e 3 dos JupyterLab notebooks para fins de compatibilidade com versões anteriores, consulte Instalação JupyterLab e extensões do Jupyter Server.

Repositórios Git

SageMaker O AI Studio vem pré-instalado com uma extensão Jupyter Git para que os usuários entrem em um repositório Git personalizadoURL, clonem em seu diretório, enviem alterações e visualizem o histórico de commits. EFS Os administradores podem configurar repositórios git sugeridos no nível do domínio para que eles apareçam como seleções suspensas para os usuários finais. Consulte Anexar repositórios Git sugeridos ao Studio para obter instruções. up-to-date

Se um repositório for privado, a extensão solicitará que o usuário insira suas credenciais no terminal usando a instalação padrão do git. Como alternativa, o usuário pode armazenar as credenciais ssh em seu EFS diretório individual para facilitar o gerenciamento.

Ambiente Conda

SageMaker Os notebooks AI Studio usam a Amazon EFS como uma camada de armazenamento persistente. Os cientistas de dados podem usar o armazenamento persistente para criar ambientes conda personalizados e usar esses ambientes para criar kernels. Esses kernels são apoiados EFS e persistem entre as reinicializações do kernel, do aplicativo ou do Studio. O Studio seleciona automaticamente todos os ambientes válidos como KernelGateway kernels.

O processo para criar um ambiente conda é simples para um cientista de dados, mas os kernels levam cerca de um minuto para serem preenchidos no seletor de kernel. Para criar um ambiente, execute o seguinte em um terminal do sistema:

```
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
```

Para obter instruções detalhadas, consulte a seção Ambientes Persist Conda para o EFS volume Studio em Quatro abordagens para gerenciar pacotes Python em notebooks Amazon Studio. SageMaker

Conclusão

Neste whitepaper, analisamos várias práticas recomendadas em áreas como modelo operacional, gerenciamento de domínio, gerenciamento de identidades, gerenciamento de permissões, gerenciamento de rede, registro, monitoramento e personalização para permitir que os administradores da plataforma configurem e SageMaker gerenciem o AI Studio Platform.

Apêndice

Comparação: multilocação

Tabela 2 — Comparação de locação múltipla

Vários domínios	Conta múltipla	Controle de acesso baseado em atributos (ABAC) em um único domínio
O isolamento de recursos é obtido usando tags. SageMaker O AI Studio marca automaticamente todos os recursos com o domínio ARN e o perfil/espaço do usuário. ARN	Cada inquilino está em sua própria conta, portanto, há isolamento absoluto de recursos.	O isolamento de recursos é obtido usando tags. Os usuários precisam gerenciar a marcação dos recursos criados paraABAC.
A lista APIs não pode ser restringida por tags. A filtragem de recursos da interface do usuário é feita em espaços compartilhados, no entanto, as API chamadas de lista feitas por meio do AWS CLI ou do Boto3 SDK listarão os recursos em toda a região.	O APIs isolamento da lista também é possível, já que os inquilinos estão em suas contas dedicadas.	A lista APIs não pode ser restringida por tags. Listar API chamadas feitas por meio do AWS CLI ou do Boto3 SDK listará recursos em toda a região.
SageMaker Os custos de computação e armazenam ento do AI Studio por locatário podem ser facilmente monitorados usando o Domain ARN como uma etiqueta de alocação de custos.	SageMaker Os custos de computação e armazenam ento do AI Studio por locatário são fáceis de monitorar com uma conta dedicada.	SageMaker Os custos de computação do AI Studio por inquilino precisam ser calculados usando tags personalizadas. SageMaker Os custos de armazenamento do AI Studio

Vários domínios	Conta múltipla	Controle de acesso baseado em atributos (ABAC) em um único domínio
		não podem ser monitorados por domínio, pois todos os locatários compartilham o mesmo EFS volume.
As cotas de serviço são definidas no nível da conta, portanto, um único inquilino ainda pode usar todos os recursos.	As cotas de serviço podem ser definidas no nível da conta para cada inquilino.	As cotas de serviço são definidas no nível da conta, portanto, um único inquilino ainda pode usar todos os recursos.
A escalabilidade para vários locatários pode ser obtida por meio da infraestrutura como código (IaC) ou do Service Catalog.	A escalabilidade para vários inquilinos envolve Organizations e a venda de várias contas.	O escalonamento precisa de uma função específica de inquilino para cada novo inquilino, e os perfis de usuário precisam ser marcados manualmente com os nomes dos inquilinos.
A colaboração entre usuários dentro de um locatário é possível por meio de espaços compartilhados.	A colaboração entre o usuário dentro de um inquilino é possível por meio de espaços compartilhados.	Todos os inquilinos terão acesso ao mesmo espaço compartilhado para colaboração.

SageMaker Backup e recuperação de domínios do AI Studio

No caso de uma EFS exclusão acidental ou quando um domínio precisar ser recriado devido a alterações na rede ou na autenticação, siga estas instruções.

Opção 1: fazer backup do EFS uso existente EC2

SageMaker Backup de domínio do Studio

1. Listar perfis de usuário e espaços no SageMaker Studio (CLI, SDK).

2. Mapeie perfis/espaços de usuário para UIDs ativado. EFS
   a. Para cada usuário na lista de users/spaces, describe the user profile/space (CLI, SDK).
   b. Mapeie o perfil/espaço do usuário para `HomeEfsFileSystemUid`.
   c. Mapeie o perfil do usuário para `UserSettings['ExecutionRole']` saber se os usuários têm perfis de execução distintas.
   d. Identifique o perfil padrão de execução do Space.
3. Crie um novo domínio e especifique o perfil de execução padrão do Space.
4. Crie perfis e espaços de usuário.
   - Para cada usuário na lista de usuários, crie um perfil de usuário (CLI, SDK) usando o mapeamento da função de execução.
5. Crie um mapeamento para o novo EFS UIDs e.
   a. Para cada usuário na lista de usuários, descreva o perfil do usuário (CLI, SDK).
   b. Mapeie o perfil do usuário para `HomeEfsFileSystemUid`.
6. Opcionalmente, exclua todos os aplicativos, perfis de usuário, espaços e, em seguida, exclua o domínio.

Backup do EFS

Para fazer backupEFS, use as seguintes instruções:

1. Inicie a EC2 instância e anexe os grupos de segurança de entrada/saída do antigo domínio do SageMaker Studio à nova EC2 instância (permita o NFS tráfego na porta 2049). TCP Consulte Connect SageMaker Studio Notebooks em Recursos Externos. VPC
2. Monte o EFS volume do SageMaker Studio na nova EC2 instância. Consulte Montagem de sistemas de EFS arquivos.
3. Copie os arquivos para o armazenamento EBS local: `>sudo cp -rp /efs /studio-backup:`
   a. Anexe os novos grupos de segurança do domínio à EC2 instância.
   b. Monte o novo EFS volume na EC2 instância.
   c. Copie os arquivos para o novo EFS volume.
   d. Para cada usuário na coleção do usuário:
      i. Crie o diretório: `mkdir new_uid`.
      ii. Copie arquivos do UID diretório antigo para o novo UID diretório.
      iii. Alterar a propriedade de todos os arquivos: `chown <new_UID>` de todos os arquivos.

Opção 2: fazer backup do existente EFS usando o S3 e a configuração do ciclo de vida

1. Consulte Migrar seu trabalho para uma instância de SageMaker notebook da Amazon com o Amazon Linux 2.
2. Crie um bucket do S3 para backup (como `>studio-backup`.
3. Liste todos os perfis de usuário com perfis de execução.
4. No domínio atual do SageMaker Studio, defina um LCC script padrão no nível do domínio.
   - NoLCC, copie tudo `/home/sagemaker-user` para o prefixo do perfil do usuário no S3 (por exemplo,`s3://studio-backup/studio-user1`).
5. Reinicie todos os aplicativos padrão do Jupyter Server (para LCC que sejam executados).
6. Exclua todos os aplicativos, perfis de usuário e domínios.
7. Crie um novo domínio do SageMaker Studio.
8. Crie novos perfis de usuário a partir da lista de perfis de usuário e perfis de execução.
9. Configure um LCC no nível do domínio:
   - NoLCC, copie tudo no prefixo do perfil do usuário no S3 para `/home/sagemaker-user`
10. Crie aplicativos padrão do Jupyter Server para todos os usuários com a LCCconfiguração (CLI, SDK).

SageMaker Acesso ao estúdio usando SAML asserção

Configuração da solução:

1. Crie um SAML aplicativo em seu IdP externo.
2. Configure o IdP externo como um provedor de identidade em. IAM
3. Crie uma função `SAMLValidator` Lambda que possa ser acessada pelo IdP (por meio de uma função URL ou Gateway). API
4. Crie uma função `GeneratePresignedUrl` Lambda e um API Gateway para acessar a função.
5. Crie uma IAM função que os usuários possam assumir para invocar o API Gateway. Essa função deve ser passada em SAML asserção como um atributo no seguinte formato:
   - Nome do atributo: https://aws.amazon.com/SAML/ Atributos/Função
   - Valores de atributo `<IdentityProviderARN>`, `<RoleARN>`

6. Atualize o endpoint SAML Assertion Consumer Service (ACS) para a `SAMLValidator` invocação. URL

SAMLcódigo de exemplo do validador:

```
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
```

Outras fontes de leitura

- Configurando ambientes de aprendizado de máquina seguros e bem governados em AWS(blog)AWS
- Configurando o Amazon SageMaker AI Studio para equipes e grupos com isolamento completo de recursos (AWS blog)
- Integração do Amazon SageMaker AI Studio com o AWS SSO Okta Universal Directory (blog)AWS
- Como configurar SAML 2.0 para federação de AWS contas (documentação do Okta)
- Crie uma plataforma de Machine Learning empresarial segura na AWS (guia técnico da AWS)
- Personalize o Amazon SageMaker AI Studio usando configurações de ciclo de vida (blog)AWS
- Trazendo sua própria imagem de contêiner personalizada para os cadernos do Amazon SageMaker AI Studio (AWS blog)
- Crie modelos de projetos de SageMaker IA personalizados — Melhores práticas (AWS blog)
- Implantação de modelo de várias contas com o Amazon SageMaker AI Pipelines (blog)AWS
- Parte 1: Como o NatWest Grupo criou uma MLOps plataforma escalável, segura e sustentável (AWS blog)
- Proteja o Amazon SageMaker AI Studio pré-assinado, URLs parte 1: infraestrutura fundamental (blog)AWS

Colaboradores

Os colaboradores deste documento incluem:

- Ram Vittal, arquiteto de soluções de ML, Amazon Web Services
- Sean Morgan, arquiteto de soluções de ML, Amazon Web Services
- Durga Sury, arquiteta de soluções de ML, Amazon Web Services

Agradecimentos especiais aos seguintes que contribuíram com ideias, revisões e perspectivas:

- Alessandro Cerè, arquiteto de soluções de IA/ML, Amazon Web Services
- Sumit Thakur, líder de produtos de SageMaker IA, Amazon Web Services
- Han Zhang, engenheiro sênior de desenvolvimento de software, Amazon Web Services
- Bhadrinath Pani, engenheiro de desenvolvimento de software, Amazon Web Services, Amazon Web Services

Revisões do documento

Para ser notificado sobre atualizações desse whitepaper, inscreva-se no feed RSS.

Alteração	Descrição	Data
Whitepaper atualizado	Links quebrados foram corrigidos e várias mudanças editoriais por toda parte.	25 de abril de 2023
Publicação inicial	Publicação do whitepaper.	19 de outubro de 2022

Avisos

Os clientes são responsáveis por fazer sua própria avaliação independente das informações contidas neste documento. Este documento: (a) é apenas para fins informativos, (b) representa as ofertas e práticas de produtos atuais da AWS, que estão sujeitas a alterações sem aviso prévio e (c) não criam nenhum compromisso ou garantia da AWS e de suas afiliadas, fornecedores ou licenciadores. Os produtos ou serviços da AWS são fornecidos “no estado em que se encontram”, sem garantias, representações ou condições de qualquer tipo, expressas ou implícitas. As responsabilidades e as obrigações da AWS com os seus clientes são controladas por contratos da AWS, e este documento não é parte, nem modifica, qualquer contrato entre a AWS e seus clientes.

© 2022 Amazon Web Services, Inc. ou suas afiliadas. Todos os direitos reservados.

AWS Glossário

Para obter a terminologia mais recente da AWS, consulte o glossário da AWS na Referência do Glossário da AWS.

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.