AWS Livro branco

SageMaker Práticas recomendadas de administração do Studio

SageMaker Práticas recomendadas de administração do Studio: AWS Livro branco

Table of Contents

Práticas recomendadas de administração do SageMaker Studio

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

Resumo

O Amazon SageMaker 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. O SageMaker 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, registro, monitoramento e personalização. As melhores práticas discutidas aqui se destinam à implantação corporativa do SageMaker Studio, 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 AWSWell-Architected Framework 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 cargas de trabalho em relação a essas melhores práticas 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 aprendizado de máquina no Nuvem AWS. Essa lente se soma às melhores práticas descritas no Well-Architected Framework.

Introdução

Ao administrar o SageMaker Studio como sua plataforma de ML, você precisa de orientação sobre as melhores práticas para tomar decisões informadas que o ajudem 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 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 proteções para várias funções de suas personas de ML.
- Planeje sua topologia de rede de nuvem privada virtual (VPC), considerando a sensibilidade da sua 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 do SageMaker 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 do Centro de Excelência (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 funções ou divisões de negócios, ou com base em diferentes linhas de produtos.
- Equipes federadas de ciência de dados — Nesse modelo, funções 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 função 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

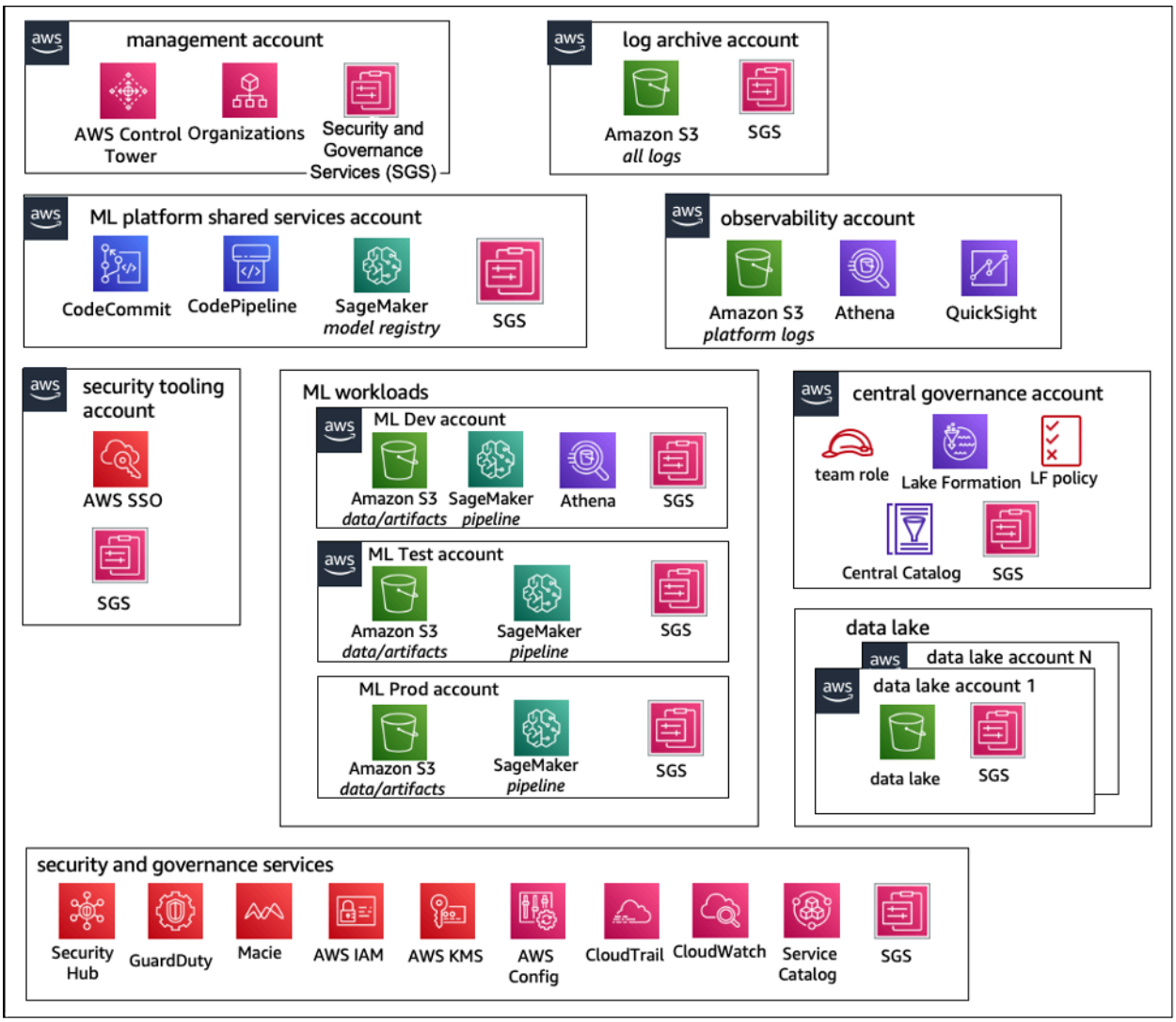
Nesta seção, apresentamos brevemente uma estrutura de conta do modelo operacional com a qual você pode começar e modificar de acordo com os requisitos operacionais da sua organização. Independentemente do modelo operacional escolhido, recomendamos implementar as seguintes práticas recomendadas comuns:

- Use AWS Control Towerpara configuração, gerenciamento e governança de suas contas.
- Centralize suas identidades com seu provedor de identidade (IdP) e o AWSIAM Identity Center com uma conta delegada do Security Tooling de administrador e habilite 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 registros 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:

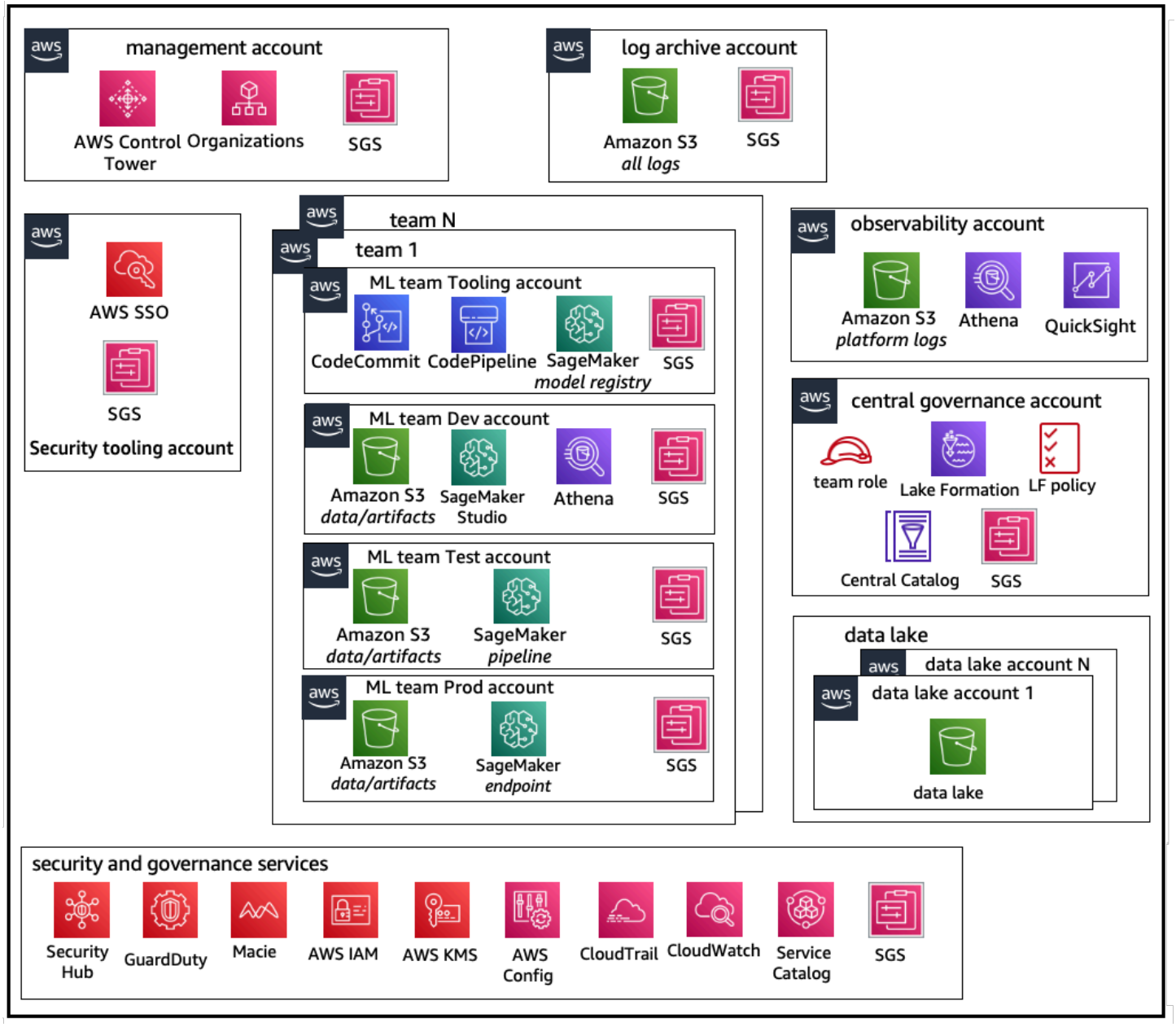
- Uma conta de ferramentas de serviços compartilhados que atende aos requisitos de Machine Learning Operations (MLOps) em todas as equipes de ciência de dados.
- Contas de desenvolvimento, teste e produção de cargas de trabalho de ML que são compartilhadas entre as equipes de ciência de dados.
- Políticas 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

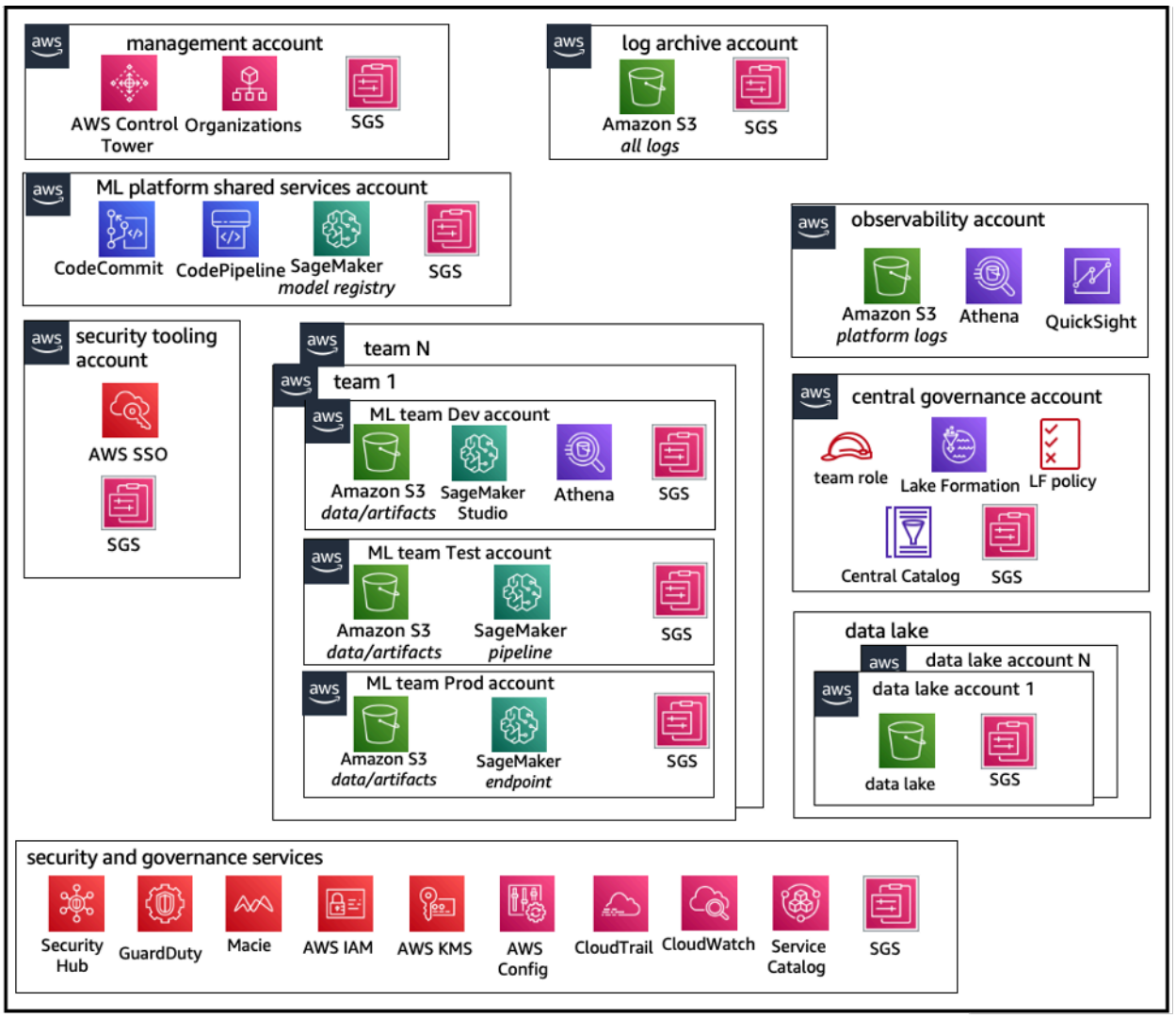
Nesse modelo, cada equipe de ML opera de forma independente para provisionar, gerenciar e governar contas e recursos de ML. No entanto, recomendamos que as equipes de ML usem uma abordagem centralizada de observabilidade e modelo de governança de dados para simplificar a governança de dados e o gerenciamento 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 equipe de ciência de dados/ML tem seu próprio conjunto de contas de carga de trabalho de desenvolvimento/teste/produção que permitem um isolamento físico robusto de seus recursos de ML e também permitem 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 inquilino 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 inquilina ou equipe.

Embora seja possível implementar várias equipes em uma instância do SageMaker Studio (como o SageMaker 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

SageMaker Studio. 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 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.

Os recursos do SageMaker criados a partir de um domínio são marcados automaticamente com o nome de recurso da Amazon (ARN) do domínio e o ARN do perfil do usuário ou do espaço 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 abas para domínios existentes no repositório do GitHub.

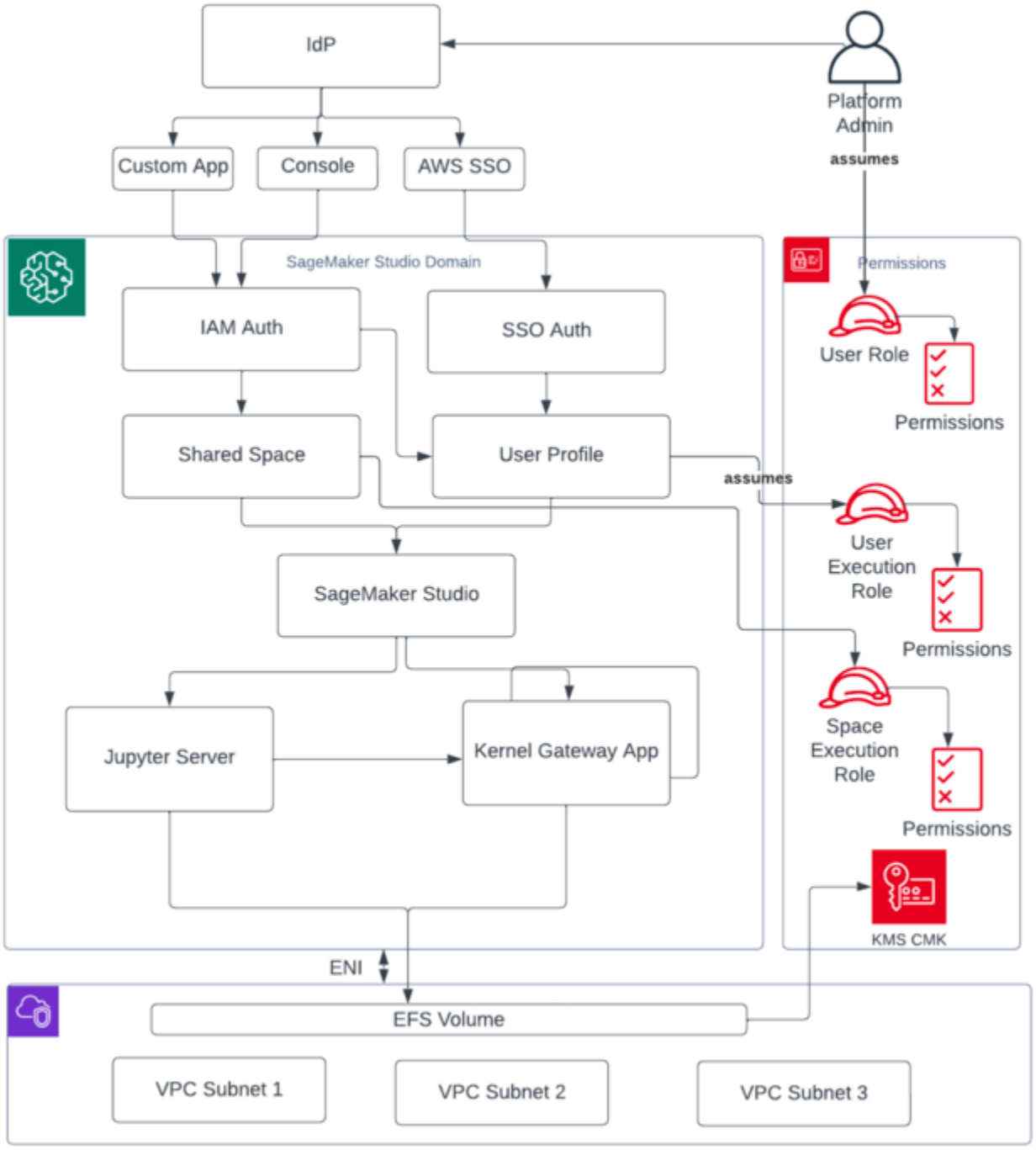
Por fim, você pode implementar uma implantação de autoatendimento dos recursos do SageMaker Studio em várias contas usando AWS Service Catalog. Para obter mais informações, consulte Gerenciar AWS Service Catalog produtos em vários Contas da AWS e Regiões da AWS.

Gerenciamento de domínio

Um domínio do Amazon SageMaker consiste em:

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

O diagrama a seguir fornece uma visão de alto nível de vários componentes que constituem um domínio do SageMakerStudio:

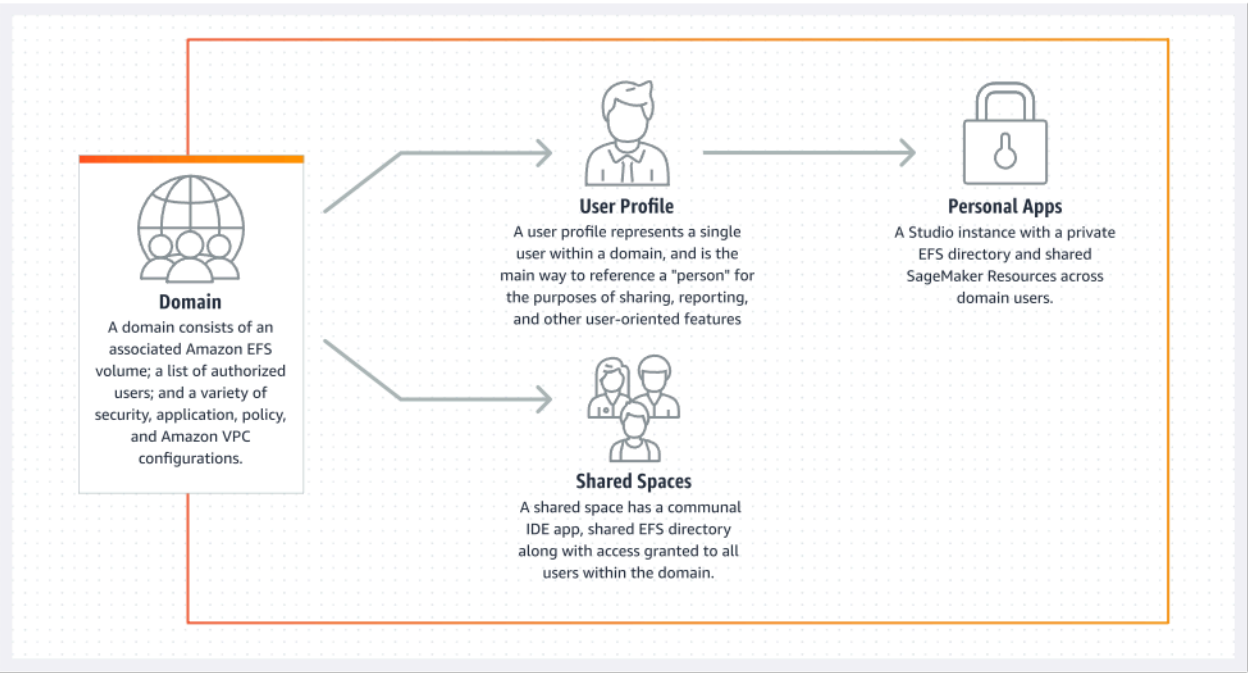


Visão de alto nível de vários componentes que constituem um domínio do SageMaker Studio

Vários domínios e espaços compartilhados

A Amazon SageMaker agora suporta a criação de vários SageMaker domínios 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 e 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 autenticação do IAM pode usar o espaço compartilhado para colaboração quase em tempo real entre os usuários. Com um espaço compartilhado, os usuários têm acesso a um diretório compartilhado do Amazon EFS 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 de usuário 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 empreendimento ou projeto de ML específico em que os membros de um único domínio exigem acesso quase em tempo real ao mesmo armazenamento de arquivos e IDE subjacentes. 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 uma função 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 a federação do IAM

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

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

Configure seu domínio para federação de autenticação única (SSO)

Para usar a federação de login único (SSO), você precisa habilitar sua conta AWS Organizationsde gerenciamento AWS IAM Identity Center na mesma região em que precisa executar o Studio. SageMaker As etapas de configuração do domínio são semelhantes às etapas de federação do IAM, 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 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 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 de configurações para um usuário individual e tem uma referência ao diretório inicial privado do Amazon Elastic File System (Amazon EFS) do usuário. Recomendamos criar um perfil de usuário para cada usuário físico do aplicativo SageMaker Studio. Cada usuário tem seu próprio diretório dedicado no Amazon EFS, e os perfis de usuário não podem ser compartilhados entre domínios na mesma conta.

Cada perfil de usuário que compartilha o domínio do SageMaker Studio recebe recursos computacionais dedicados (como instâncias do SageMaker Amazon Elastic Compute Cloud (Amazon EC2)) 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 AWS conta 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 Studio para um usuário acessando a URL pré-assinada ou fazendo login usando o AWS IAM iDC, o aplicativo Jupyter Server é lançado na instância VPC gerenciada pelo serviço SageMaker . Cada usuário obtém seu próprio aplicativo Jupyter Server dedicado em um aplicativo privado. Por padrão, o aplicativo Jupyter Server para notebooks SageMaker 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 API ou da interface do SageMaker Studio e é executado no tipo de instância escolhido. Esse aplicativo pode ser executado usando uma das imagens integradas do SageMaker 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 dentro do mesmo SageMaker aplicativo Studio Image/Kernel Gateway. Os usuários também podem executar até quatro aplicativos ou imagens do Kernel Gateway na mesma instância física, cada um isolado por seu contêiner/imagem.

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 Studio e outro notebook usando a TensorFlow imagem incorporada, 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 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 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 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 (Amazon EFS) é criado para uso por todos os usuários dentro do domínio. Cada perfil de usuário recebe um diretório inicial privado dentro do volume do Amazon EFS para armazenar os notebooks, GitHub repositórios e arquivos de dados do usuário. Cada espaço dentro de um domínio recebe um diretório privado dentro do volume do Amazon EFS 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 Studio cria uma ID de usuário global exclusiva para cada perfil ou espaço de usuário e a aplica como uma ID de usuário/grupo da Portable Operating System Interface (POSIX) para o diretório inicial do usuário no EFS, o que impede que outros usuários/espaços acessem seus dados.

Backup e recuperação

Um volume EFS existente não pode ser anexado a um novo SageMaker domínio. Em uma configuração de produção, certifique-se de que o volume do Amazon EFS seja copiado (para outro

volume do EFS ou para o Amazon Simple Storage Service (Amazon S3)). Se um volume EFS for excluído acidentalmente, o administrador precisará desmontar e recriar o domínio do SageMaker Studio. O processo é o seguinte:

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

1. Crie um novo domínio do SageMaker 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 Studio.

Para obter instruções detalhadas, consulte a seção Backup e recuperação de domínio do SageMaker Studio no apêndice.

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

Volume do Amazon EBS

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

Protegendo o acesso ao URL pré-assinado

Quando um usuário do SageMaker Studio abre o link do notebook, o SageMaker Studio valida a política de IAM do usuário federado para autorizar o acesso e gera e resolve a URL pré-assinada para o usuário. Como o SageMaker console é executado em um domínio da Internet, esse URL

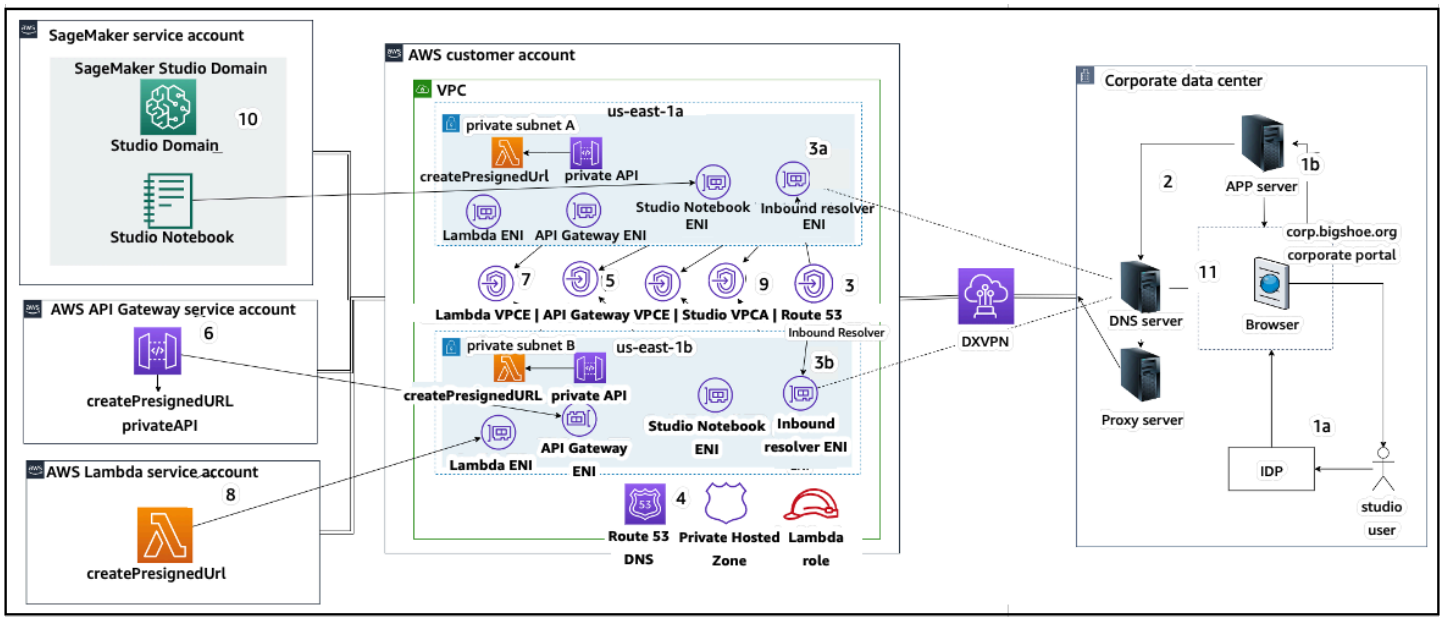
gerado e pré-assinado fica visível na sessão do navegador. Isso representa um vetor de ameaça indesejado para roubo de dados e para a obtenção de acesso aos dados do cliente quando os controles de acesso adequados não são aplicados.

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

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

Quando você acessa os notebooks do SageMaker Studio pelo SageMaker console, a única opção disponível é usar a validação de IP do cliente com a condição `aws:sourceIp` de política do IAM. 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 `aws:sourceIp` condição.

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



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

Para step-by-step obter orientação sobre a configuração da arquitetura anterior, consulte URLs pré-assinados seguros do Amazon SageMaker Studio, Parte 1: Infraestrutura básica.

SageMaker cotas e limites de domínio

- SageMaker A federação de SSO do domínio 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.
- A configuração da VPC e da sub-rede não pode ser alterada após a criação do domínio. No entanto, você pode criar um novo domínio com uma configuração diferente de VPC e sub-rede.
- O acesso ao domínio não pode ser alternado entre os modos IAM e SSO 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 proteções 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 acessam o Studio. SageMaker Primeiro, descreveremos brevemente como usuários, grupos e perfis são mapeados e como funciona a federação de usuários.

Usuários, grupos e perfil

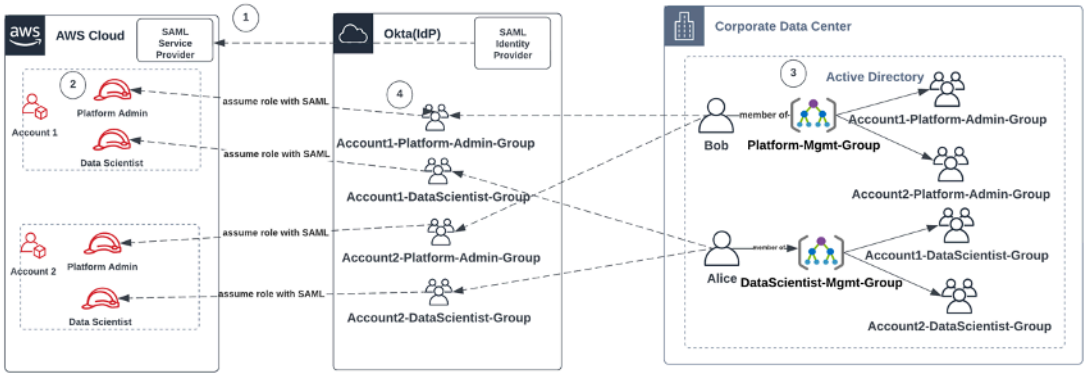
Em AWS, as permissões de recursos são gerenciadas usando usuários, grupos e funções. Os clientes podem gerenciar seus usuários e grupos por meio do IAM ou em um diretório corporativo, como o Active Directory (AD), habilitado por meio de um IdP externo, como o Okta, que permite autenticar os usuários em vários aplicativos executados na nuvem e on-premises.

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 funções Contas da AWS e tenham acesso a funções específicas usando o SSO com a linguagem de marcação de asserção de segurança (SAML). Os administradores do IdP podem 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 funções do IAM. As principais etapas do fluxo de trabalho são as seguintes:



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

1. Em AWS, configure a integração SAML para 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). SageMaker

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

Federação de usuários

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

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

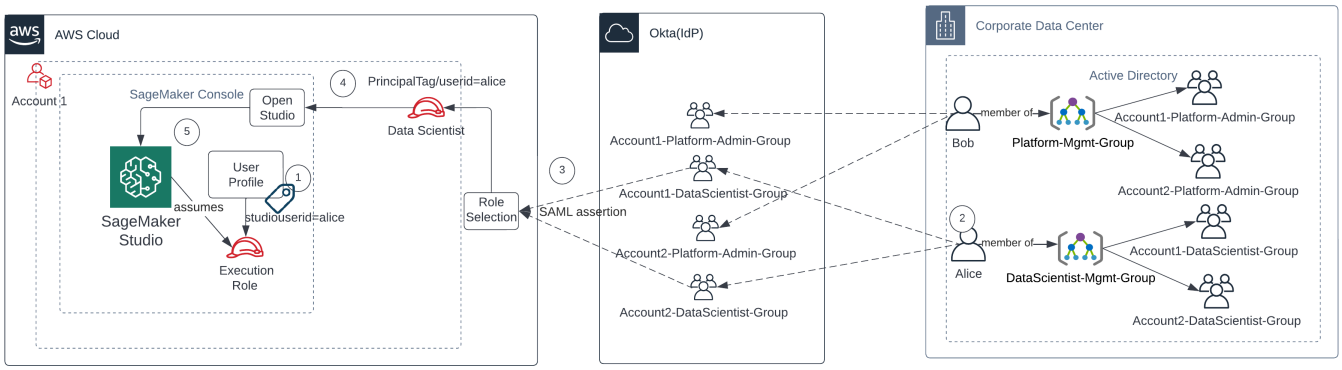
usuários do IAM

Para usuários do IAM, o administrador cria perfis de usuário do SageMaker Studio para cada usuário e associa o perfil do usuário a uma função do IAM que permite as ações necessárias que o usuário precisa realizar dentro do Studio. Para impedir que um AWS usuário acesse somente seu perfil de usuário do SageMaker Studio, o administrador deve marcar o perfil de usuário do SageMaker Studio e anexar uma política do IAM 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 IAM ou federação de contas

O método de Conta da AWS federação permite que os clientes se federem no SageMaker console a partir de seu IdP SAML, como o Okta. Para impedir que os usuários acessem somente seu perfil de usuário, o administrador deve marcar o perfil de usuário do SageMaker 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 Studio.



Acessando o SageMaker Studio no modo de federação do IAM

1. O perfil de usuário do Alice SageMaker 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 declaração SAML com as duas perfis (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 SageMaker console 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 Studio selecionada. Se a tag de perfil for válida, a instância do aplicativo SageMaker Studio será iniciada, assumindo a função de execução.

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

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

Configure um processo de automação que escute o evento de `SageMaker-Account1-Group` associação e use AWS APIs para criar a função, as políticas, as tags e o perfil de usuário do SageMaker 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 Studio acessem outros perfis de usuário.

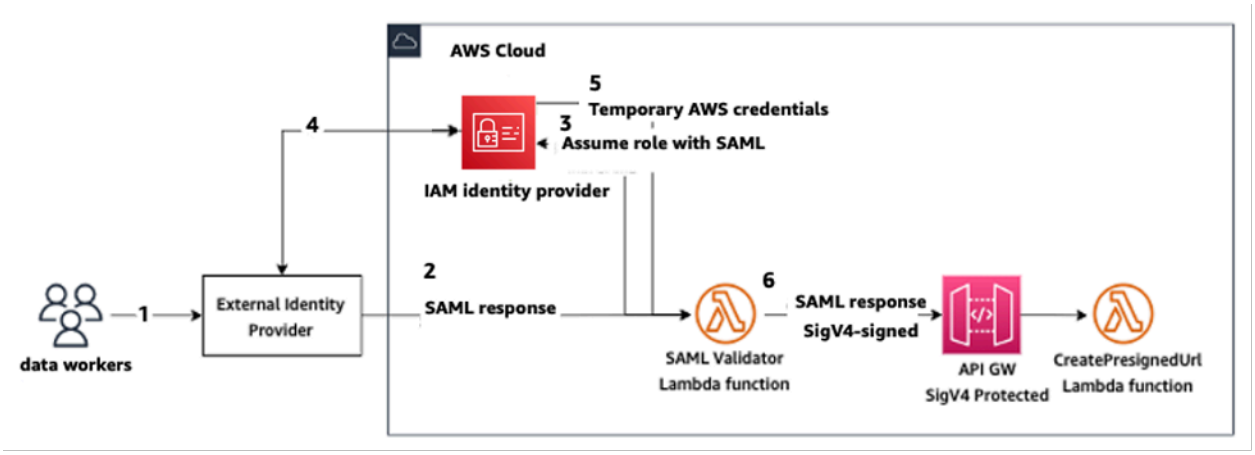
Autenticação SAML usando AWS Lambda

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

A vantagem dessa solução é que a função Lambda pode personalizar a lógica para acesso ao SageMaker 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 Studio analisando os atributos SAML.
- Personalize o perfil do usuário adicionando Life Cycle Configuration (LCC) e adicionando tags.

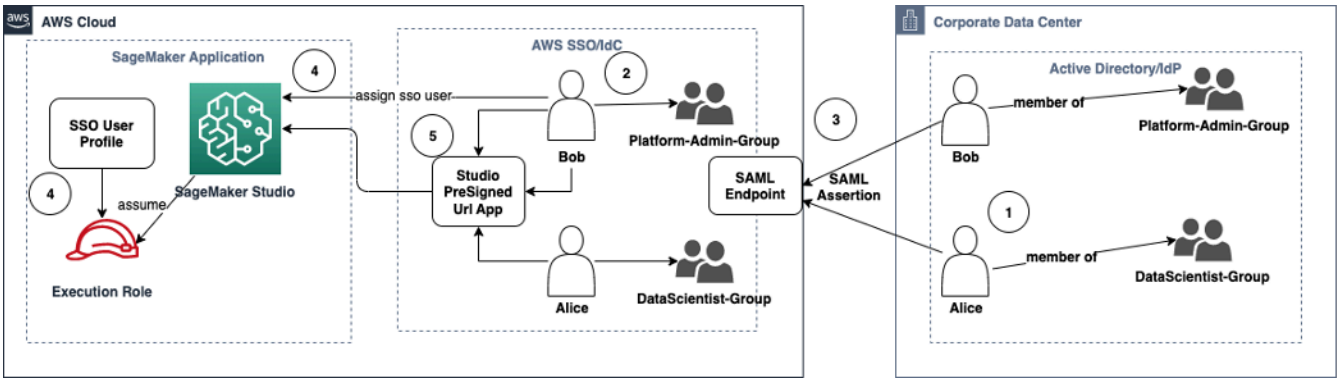
Em resumo, essa solução exporá o SageMaker Studio como um aplicativo SAML2.0 com lógica personalizada para autenticação e autorização. Consulte a seção do apêndice Acesso ao SageMaker Studio usando a declaração SAML para obter detalhes da implementação.



Acessando o SageMaker Studio usando um aplicativo SAML personalizado

Federação AWS IAM IdC

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



Acessando o SageMaker Studio no modo IAM 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 Identity Provider (IdP) são sincronizados com o AWS IAM Identity Center e estão disponíveis como usuários e grupos de login único para tarefas, respectivamente.
3. O IdP publica uma declaração SAML no endpoint SAML do AWS IdC.
4. No SageMaker Studio, o usuário do iDC é atribuído ao aplicativo SageMaker Studio. Essa tarefa pode ser feita usando o iDC Group e o SageMaker Studio será aplicado em cada nível de usuário

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

5. O usuário acessa o aplicativo SageMaker Studio usando o URL seguro pré-assinado hospedado como um aplicativo em nuvem do iDC. SageMaker O 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 IU do SageMaker Studio diretamente, use o modo de login único com o AWS IAM iDC.
2. Se você quiser que seus usuários não acessem AWS Management Console e visualizem a interface do usuário do SageMaker Studio diretamente no modo IAM, você pode fazer isso usando uma função Lambda no back-end para gerar uma URL pré-assinada para o perfil do usuário e redirecioná-los para a interface do usuário do 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 a UpdateUserProfileAPI.
5. Se você quiser restringir o acesso ao SageMaker Studio UI no modo IAM (usando o URL pré-assinado gerado) a um VPC endpoint, sem atravessar a Internet, você pode usar um resolvedor de DNS personalizado. Consulte a postagem do blog sobre URLs pré-assinados do Secure Amazon SageMaker Studio, Parte 1: Infraestrutura fundamental.

Gerenciamento de permissões

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

Perfis 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 o SageMaker é um serviço gerenciado, você também precisa considerar os princípios de serviço, que AWS são serviços que podem fazer chamadas de API em nome do usuário. O diagrama a seguir ilustra as diferentes funções do IAM que você pode querer criar, correspondendo às diferentes personas na organização.

Funções do SageMaker IAM

Essas funções são descritas em detalhes, junto com alguns exemplos de permissões de IAM específicas que elas precisarão.

- Função de usuário do administrador de ML — Esse é um 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 repositórios Amazon ECR 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 Studio e fornecer produtos do Service Catalog, como projetos personalizados e modelos do Amazon EMR.

  Como esse diretor não executará trabalhos de treinamento, por exemplo, ele não precisa de permissões para iniciar trabalhos de treinamento ou processamento do SageMaker. Se eles estiverem usando a infraestrutura como modelos de código, como CloudFormation ou 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 leitura ao SageMaker usando o AWS Management Console.

Essa função de usuário também precisará de determinadas permissões do EC2 para iniciar o domínio dentro de uma VPC privada, permissões KMS para criptografar o volume EFS, bem como permissões para criar uma função vinculada ao serviço para Studio (`iam:CreateServiceLinkedRole`). Descreveremos essas permissões granulares posteriormente no documento.

- Função de usuário de Cientista de Dados — Esse princípio é o usuário que faz login no SageMaker 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 Studio, e o restante das políticas pode ser gerenciado pela função de serviço de execução do SageMaker.
- Função do serviço de execução do SageMaker — Como o SageMaker é um serviço gerenciado, ele lança trabalhos em nome do usuário. Essa função geralmente é a mais ampla em termos de permissões permitidas, porque muitos clientes optam por usar uma única função de execução para executar trabalhos de treinamento, trabalhos de processamento ou trabalhos de hospedagem de modelos. Embora essa seja uma maneira fácil de começar, já que os clientes amadurecem em sua jornada, eles geralmente dividem a função de execução do notebook em funções separadas para diferentes ações de API, especialmente ao executar essas tarefas em ambientes implantados.

  Você associa uma função ao domínio do SageMaker Studio após a 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 suas funções de trabalho), você também pode associar uma função do IAM 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 também a função de execução do domínio do SageMakerStudio ao perfil do usuário.

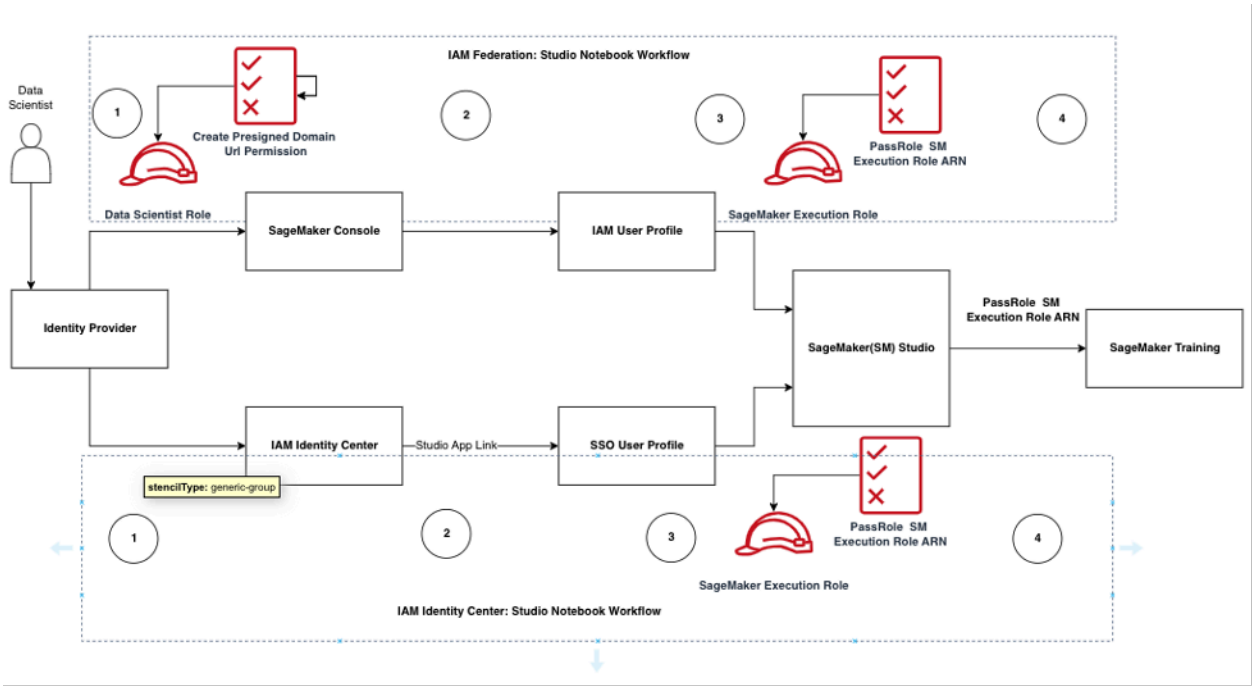
  Nos casos em que vários cientistas de dados e engenheiros de ML trabalham juntos em um projeto e precisam de um modelo de permissão compartilhado para acessar recursos, recomendamos que você crie uma função de execução de serviço do SageMaker em nível de equipe para compartilhar as permissões do IAM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

Fluxo de trabalho de autorização do SageMaker Studio Notebook

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

- Federação do IAM
- IAM Identity Center

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

Federação do IAM: 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 do SageMaker. Essa função de federação tem permissão de `iam:PassRole` API na função de execução do SageMaker 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 usuário do Studio IAM que está associado à função de execução do SageMaker.
3. O serviço SageMaker Studio IDE é iniciado, assumindo as permissões da função de execução do SageMaker no perfil do usuário. Essa função tem permissão de `iam:PassRole` API na função de execução do SageMaker para passar o ARN da função para o serviço de treinamento do SageMaker.
4. Quando o cientista de dados inicia o trabalho de treinamento no(s) nó(s) de computação remota, o ARN da função de execução do SageMaker é passado para o serviço de treinamento do SageMaker. Isso cria uma nova sessão de função com esse 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 o ARN dessa função ao chamar a API de treinamento.

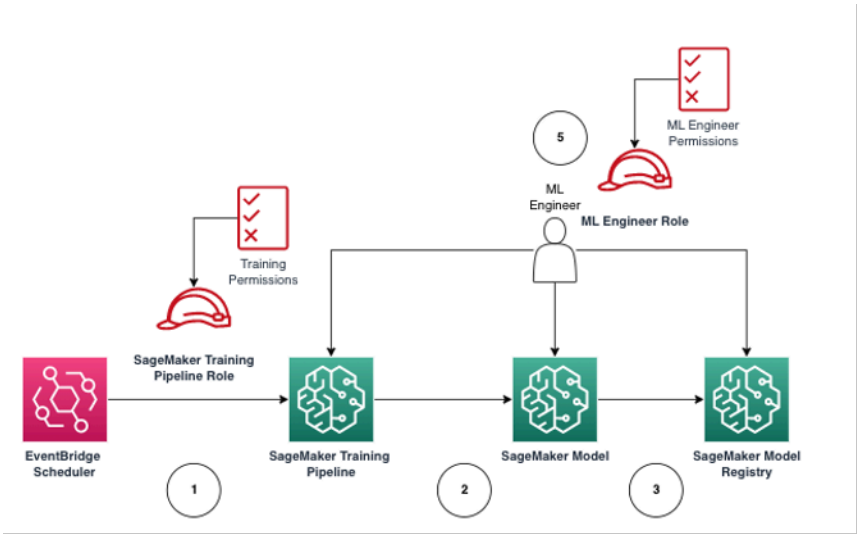
Centro de identidade do IAM: fluxo de trabalho do SageMaker Studio Notebook

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 Studio que foi criado a partir de seu perfil de usuário do iDC, que está associado à função de execução do SageMaker.
3. O serviço SageMaker Studio IDE é iniciado, assumindo as permissões da função de execução do SageMaker no perfil do usuário. Essa função tem permissão de `iam:PassRole` API na função de execução do SageMaker para passar o ARN da função para o serviço de treinamento do SageMaker.
4. Quando o cientista de dados inicia o trabalho de treinamento em nós de computação remotos, o ARN da função de execução do SageMaker é passado para o serviço de treinamento do SageMaker. O ARN da função de execução cria uma nova sessão de função com esse 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 passar o ARN dessa função ao chamar a API de treinamento.

Ambiente implantado: fluxo de trabalho de treinamento do SageMaker

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 notebooks do SageMaker Studio. Esta seção discute como as funções do IAM funcionam com o pipeline de treinamento do SageMaker no ambiente implantado.



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

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

Permissões de dados

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

Acesso a dados do AWS Lake Formation

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

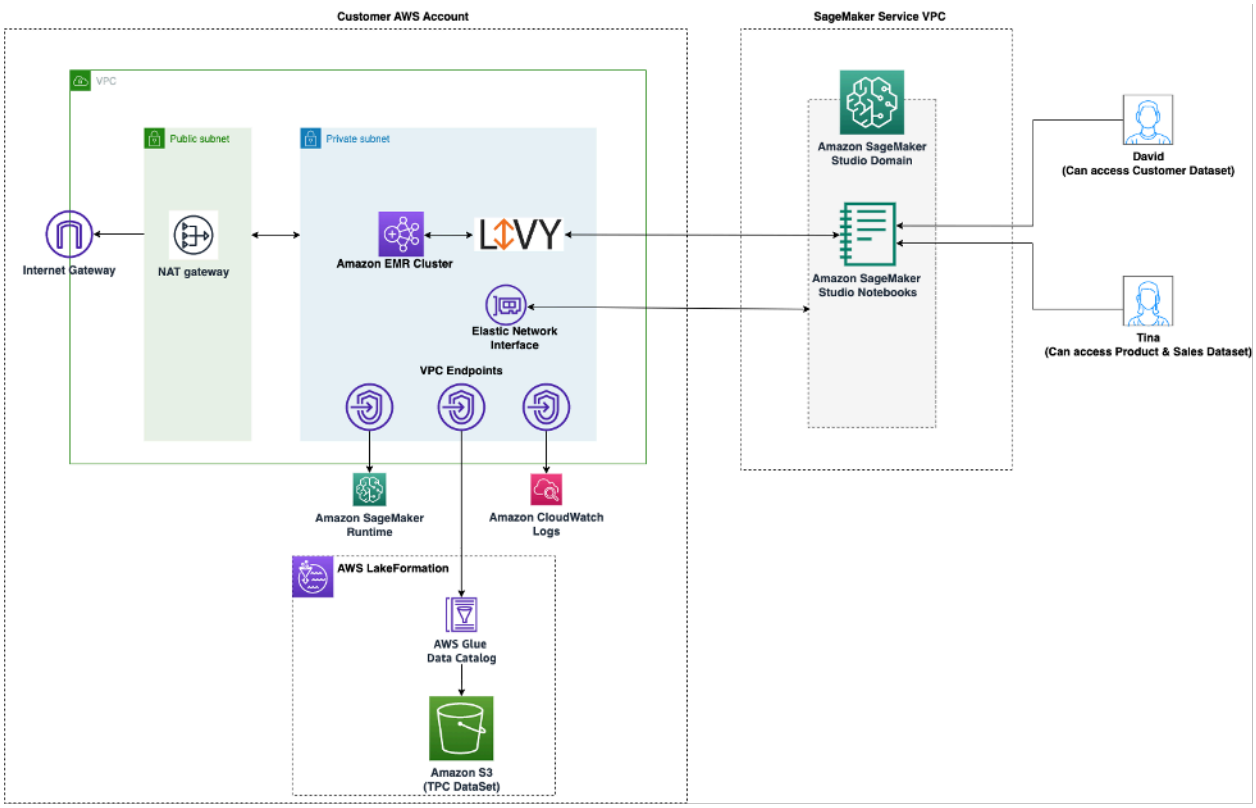
administradores podem mascarar colunas confidenciais para alguns usuários e, ao mesmo tempo, permitir consultas na mesma tabela subjacente.

Para utilizar o Lake Formation do SageMaker Studio, os administradores podem registrar as funções de execução do SageMaker IAM 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 Studio:

1. Em um notebook do SageMaker 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 AWS SDK for Pandas (anteriormente conhecido 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 Studio com o Amazon EMR para ler e gravar dados em grande escala. Com o uso das funções de tempo de execução do Apache Livy e do Amazon EMR, o SageMaker Studio criou uma conectividade nativa que permite que você transmita sua função IAM de execução do SageMaker (ou outra função autorizada) para um cluster do Amazon EMR para acesso e processamento de dados. Consulte Connect to an Amazon EMR Cluster from Studio para obter instruções atualizadas.



Arquitetura para acessar dados gerenciados pelo Lake Formation do SageMaker Studio

3. Use a conectividade nativa do SageMaker Studio em sessões interativas AWS Glue para ler e gravar dados em grande escala. Os notebooks do SageMaker 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 execução do SageMaker ou outras funções autorizadas do IAM. Consulte Preparar dados usando sessões AWS Glue interativas para obter mais informações.

Guardrails comuns

Esta seção discute as barreiras de proteção mais usadas para aplicar a governança em seus recursos de ML usando políticas de IAM, políticas de recursos, políticas de endpoint de VPC 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 padrão do Jupyter Server que hospeda o SageMaker 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"
                     ]
                 }
             }
         }
     ]
 }
```

Limitar domínios não compatíveis do SageMaker Studio

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

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

Limitar o lançamento de imagens não autorizadas do SageMaker

A política a seguir impede que um usuário inicie uma imagem não autorizada do SageMaker 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 dos endpoints VPC do SageMaker

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

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

Limitar o acesso ao notebook SageMaker Studio a um intervalo de IP limitado

As empresas geralmente limitam o acesso ao SageMaker Studio a determinados intervalos de IP corporativos permitidos. A política do IAM 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 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 Studio com a chave de tag`studiouserid`. O principal (usuário ou função 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 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 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 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 (compatível com CLI e SDK), considere adicionar esta 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 Studio

No SageMaker Studio, o notebook é executado em um contêiner Docker que, por padrão, não tem acesso root à instância host. Da mesma forma, além do usuário run-as padrão, todos os outros intervalos de IDs de usuário dentro do contêiner são remapeados como IDs de usuário sem privilégios na própria instância host. 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 root no Dockerfile. Se você criar o usuário como root ou não root, você precisa garantir que o UID/GID do usuário seja idêntico ao UID/GID no AppImageConfig do aplicativo personalizado, o que cria a configuração para o SageMaker executar um aplicativo usando a imagem personalizada. Por exemplo, se seu Dockerfile for criado para um usuário não root, 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 valores aceitáveis de UID/GID para imagens personalizadas são 0/0 e 1000/100 para imagens do Studio. Para exemplos de criação de imagens personalizadas e as `AppImageConfig` configurações associadas, consulte este repositório do Github.

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

Gerenciamento de rede

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

Planejamento da rede VPC

As sub-redes VPC do cliente associadas ao domínio SageMaker Studio devem ser criadas com o intervalo apropriado de roteamento sem classe entre domínios (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 e AWS os serviços participantes injetam interfaces de rede elástica (ENI) na sub-rede VPC do cliente para os seguintes casos de uso:

- O Amazon EFS injeta uma ENI para um destino de montagem do EFS para o SageMaker domínio (um IP por sub-rede/zona de disponibilidade anexada ao domínio). SageMaker
- SageMaker O Studio injeta uma 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 `ml.t3.medium` instância), o Studio injeta dois endereços IP.
  - Se um perfil de usuário executa um aplicativo de servidor Jupyter padrão (uma instância de “sistema”), um aplicativo de Tensorflow GPU (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.
- Uma ENI para cada VPC endpoint em todas as sub-redes/zonas de disponibilidade da VPC do domínio é injetada (quatro IPs para endpoints de VPC; ~ seis IPs para endpoints de SageMaker VPC de serviços participantes, como S3, ECR e.) CloudWatch

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

> **Note**
>
> As configurações de VPC do SageMaker Studio, como sub-redes e tráfego somente de VPC, não são repassadas automaticamente para os trabalhos de treinamento/processamento criados no Studio. SageMaker O usuário precisa definir as configurações de VPC e o isolamento de rede conforme necessário ao chamar as 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 SageMaker domínio esteja configurado no modo de tráfego somente para VPC. Existem endpoints de VPC configurados, como SageMaker API, SageMaker 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 configuração de VPC em uma `ml.m5.4xlarge` instância.

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

Tabela 1 — ENIs injetados na VPC do cliente para um cenário de experimentação

Entidade	Destino	ENI injetado	Observações	Nível
Alvo de montagem do EFS	Sub-redes VPC	Três	Três AZS/sub-redes	Domínio
Endpoints da VPC	Sub-redes VPC	30	Três AZS/sub-redes com 10 VPCE cada	Domínio

Entidade	Destino	ENI injetado	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 o EFA for usado	Usuário

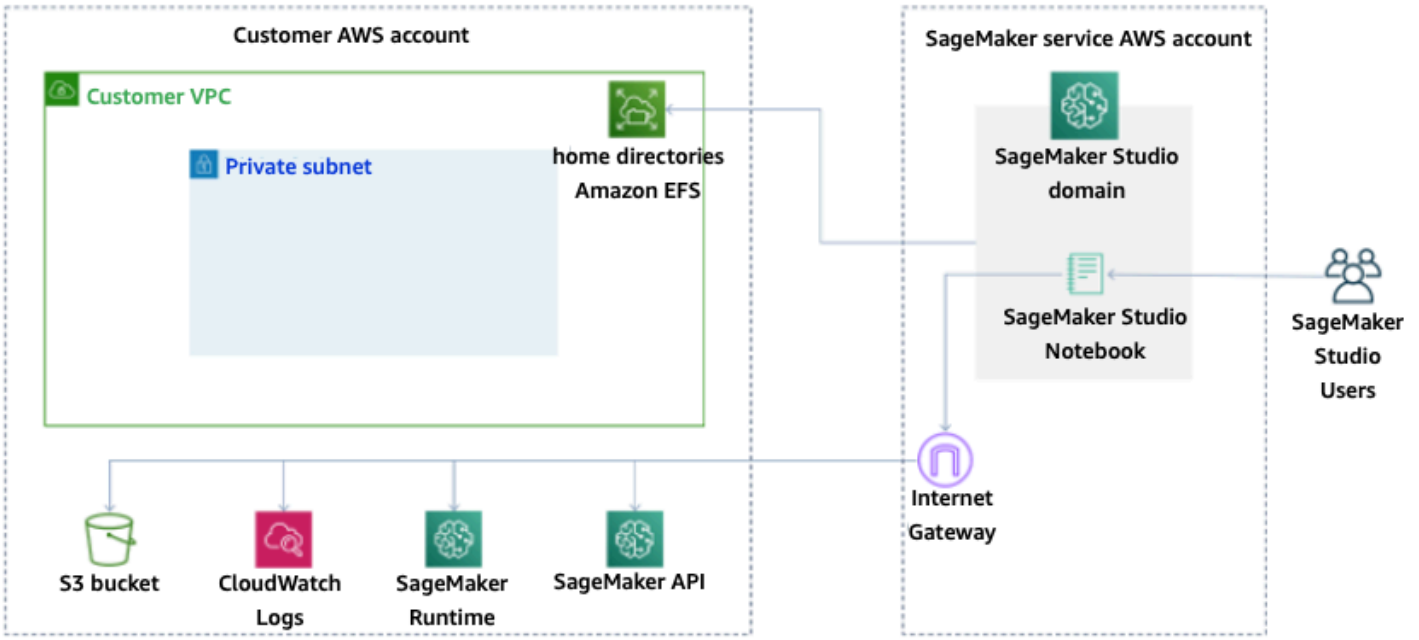
Nesse cenário, há um total de 38 IPs consumidos na VPC do cliente, em que 33 IPs são compartilhados entre usuários no nível do domínio e cinco IPs são consumidos no nível do usuário. Se você tiver 100 usuários com perfis de usuário semelhantes nesse domínio realizando essas atividades simultaneamente,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

Opções de rede de VPC

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

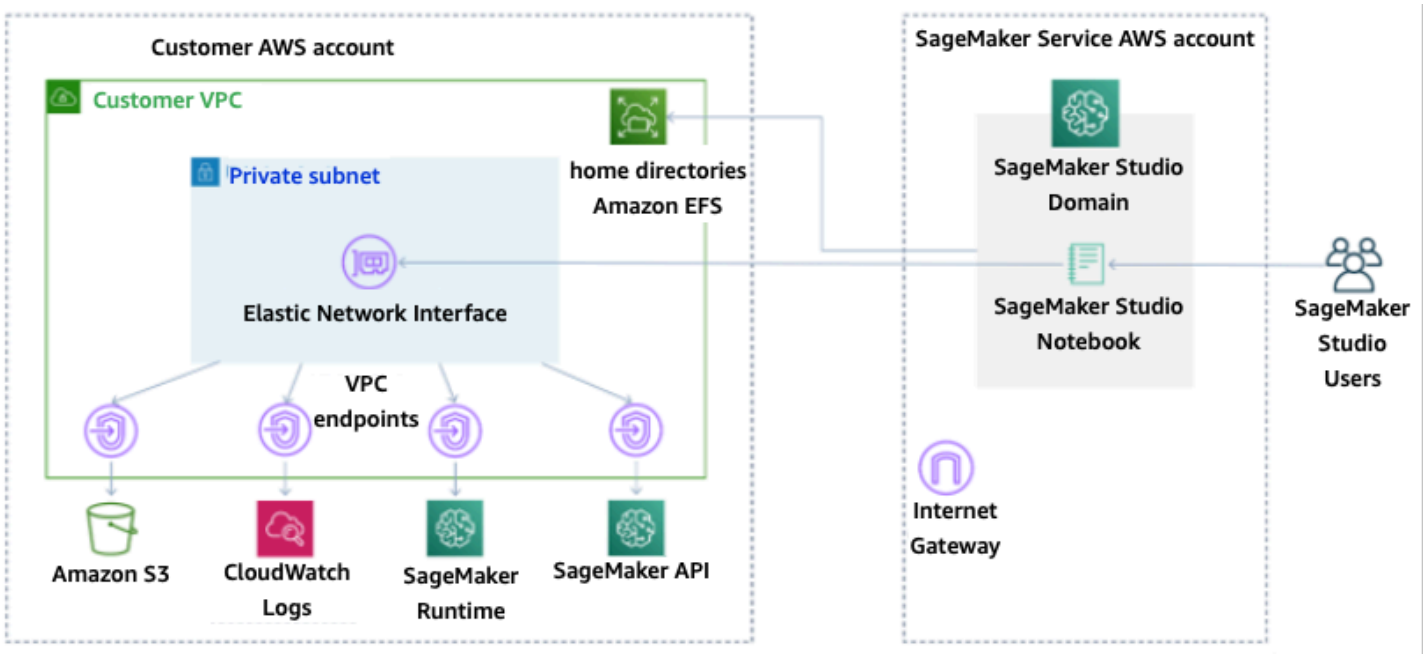
- Somente internet pública
- Somente VPC

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



Modo padrão: acesso à Internet via conta SageMaker de serviço

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



Modo somente VPC: sem acesso à Internet por meio SageMaker da conta de serviço

Para um domínio configurado somente no modo VPC, 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 configuração de VPC e modo de internet. Para obter mais detalhes sobre a configuração da rede VPC, consulte Connect SageMaker Studio Notebooks in a VPC to External Resources.

Limitações

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

O SageMaker Studio fornece vários recursos para proteger dados em repouso e em trânsito. No entanto, conforme descrito no AWSmodelo de Responsabilidade 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 cadernos do SageMaker Studio junto com seus dados de criação de modelos e artefatos de modelo, o SageMaker criptografa os cadernos, bem como a saída das tarefas de treinamento e transformação em lote. O SageMaker os criptografa por padrão, usando a chave gerenciada AWSpara 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 do SageMaker para que ela possa ser compartilhada para acesso entre contas.

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

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

Criptografia em repouso com AWS KMS

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

Proteger dados em trânsito

O SageMaker Studio garante que os artefatos de modelos de ML e outros artefatos de sistema sejam criptografados em trânsito e em repouso. As solicitações para a API e o console do SageMaker Studio são efetuadas em 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 deep learning distribuídos.

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

Guardrails de proteção de dados

Criptografe volumes de hospedagem do SageMaker em repouso

Use a política a seguir para impor a criptografia durante a hospedagem de um endpoint do SageMaker 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 os dados enviados ao seu endpoint do SageMaker e os armazena em um bucket do 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 Studio

Aplique criptografia ao volume de armazenamento anexado ao domínio do Studio. Essa política exige que o usuário forneça uma 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 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 volume EFS anexado com uma AWS KMS chave personalizada.
- Você não pode atualizar tarefas de treinamento/processamento ou configurações de endpoint com chaves KMS depois de criadas.

Registro e monitoramento

Para ajudar a depurar os trabalhos de compilação, trabalhos de processamento, os trabalhos de treinamento, os endpoints, os trabalhos de transformação, as instâncias de blocos de anotações e as configurações de ciclo de vida de instâncias de bloco de anotações, tudo o que for enviado ao ou ao por um contêiner de algoritmo, um contêiner de modelo ou uma configuração de ciclo de vida de instância de bloco de anotações também é enviado ao Amazon CloudWatch Logs. Você pode monitorar o SageMaker Studio usando o 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.

Registro em log com o 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, o SageMaker 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 para o CloudWatch sem usar a Internet pública. Também é possível definir alarmes que observam determinados limites e enviam notificações ou realizam ações quando esses limites são atingidos. Para obter mais informações, veja o Guia do usuário do Amazon CloudWatch.

O SageMaker cria um único grupo de registros para o Studio, em. `/aws/sagemaker/studio` Cada perfil de usuário e aplicativo tem seu próprio fluxo de registros nesse grupo de registros, e os scripts de configuração do ciclo de vida também têm seu próprio fluxo de registros. 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 registros:

```
/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 o SageMaker envie registros para o CloudWatch em seu nome, o chamador das APIs de trabalho de treinamento/processamento/transformação 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 serviço CloudWatch 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 nome de recurso da Amazon (ARN) específico para o log do CloudWatch que você deseja criptografar. Aqui, mostramos que você pode usar essa chave para criptografar todos os registros em uma conta para simplificar. Além disso, endpoints de treinamento, processamento e modelagem publicam métricas sobre a utilização da CPU e da memória da instância, latência de invocação de hospedagem e assim por diante. Você também pode configurar o Amazon SNS para notificar os administradores sobre eventos quando determinados limites forem ultrapassados. O consumidor das APIs de treinamento e processamento 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 todas as suas APIs com. AWS CloudTrail Por padrão, todas as APIs do SageMaker são registradas com. AWS CloudTrail Não são necessárias permissões adicionais do IAM para ativar o CloudTrail.

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

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

- Se a solicitação foi feita com credenciais de usuário 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 serviço da AWS. Para ver um exemplo de evento, consulte a documentação Log SageMaker API Calls with CloudTrail.

Por padrão, o CloudTrail registra o nome da função de execução do Studio do perfil do usuário como o identificador de cada evento. Isso funciona se cada usuário tiver sua própria função 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 o CloudTrail. Consulte Monitoramento do acesso aos recursos do usuário no Amazon SageMaker Studio para ativar o recurso`sourceIdentity`. Em um espaço compartilhado, todas as ações se referem ao ARN do espaço como fonte, e você não pode fazer auditoria`sourceIdentity`.

Atribuição de custos

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

Marcação automática

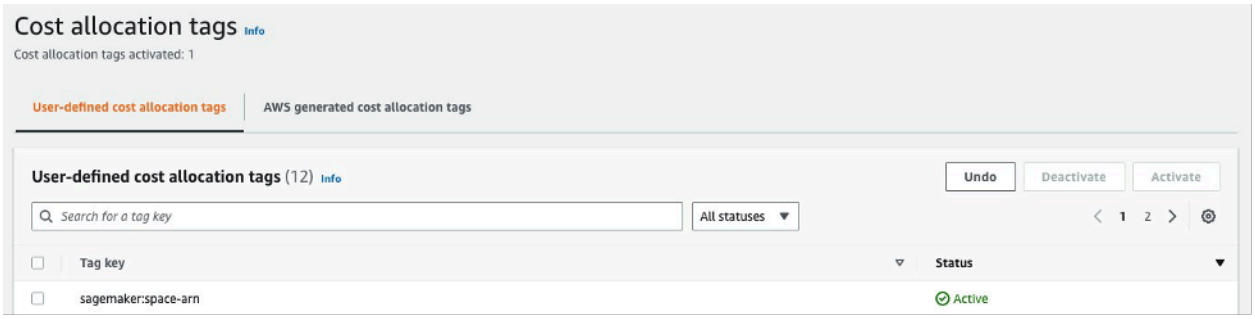
Agora, o SageMaker Studio marca automaticamente novos recursos do SageMaker, como trabalhos de treinamento, trabalhos de processamento e aplicativos do kernel, com seus respectivos `sagemaker:domain-arn`. Em um nível mais granular, o SageMaker também marca o recurso com `sagemaker:user-profile-arn` ou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

Os volumes EFS do domínio SageMaker são marcados com uma chave nomeada `ManagedByAmazonSageMakerResource` com o valor do ARN do domínio. Eles não têm tags granulares para entender o uso do espaço em um nível por usuário. Anexe o volume do EFS a uma instância do EC2 para um monitoramento personalizado.

Monitoramento de custos

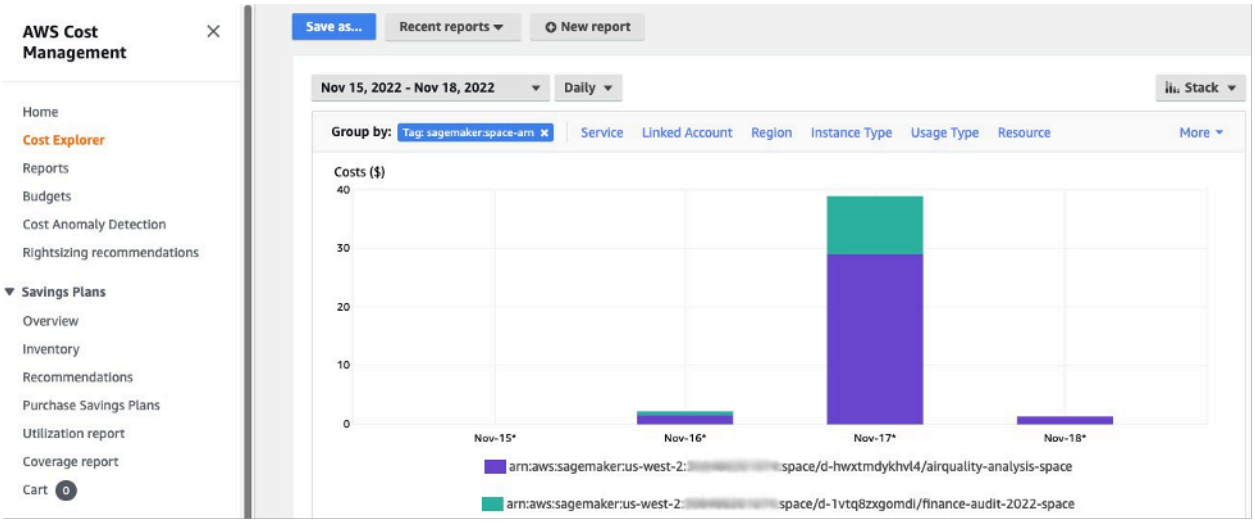
As tags automatizadas permitem que os administradores rastreiem, relatem e monitorem seus gastos com ML por meio de soluções prontas para uso, como AWS Cost Explorere AWS Budgets, bem como soluções personalizadas criadas com base nos dados dos AWSRelatórios de Custos e Uso (CURs).

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



ARN de espaço 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 Studio é integrado, o SageMaker cria um volume EFS para o domínio. Os custos de armazenamento desse volume do EFS 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 a definição de preço do Amazon SageMaker para ver uma análise detalhada dos custos.

Os administradores podem controlar os custos de computação especificando a lista de instâncias que um usuário pode criar, usando as políticas do IAM, conforme mencionado na seção Guardrails comuns. Além disso, recomendamos que os clientes usem a extensão de desligamento automático do SageMaker 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 Studio, como iniciar um novo notebook do SageMaker Studio. Você pode usar esses scripts de shell para automatizar a personalização de seus ambientes do SageMaker 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 Studio usando configurações de ciclo de vida.

Imagens personalizadas para notebooks do SageMaker Studio

Os notebooks Studio vêm com um conjunto de imagens pré-criadas, que consistem no SDK do Amazon SageMaker Python e na versão mais recente do runtime ou kernel do IPython. Com esse recurso, você pode trazer suas próprias imagens personalizadas para os cadernos do Amazon SageMaker. 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 populares de ML, como TensorFlow, MXNet, PyTorch ou outras.
- Traga códigos ou algoritmos personalizados desenvolvidos localmente para os notebooks do SageMaker Studio para acelerar a iteração e o treinamento de modelos.
- Acesso a data lakes ou armazenamentos de dados locais por meio de APIs. Os administradores precisam incluir os drivers correspondentes na imagem.
- Acesso a um tempo de execução interna (também chamado de kernel), diferente do 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 personalizada do SageMaker.

extensões do JupyterLab

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

O JupyterLab 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 Studio, abra o terminal do sistema a partir do inicializador do Studio e execute os comandos a seguir. 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 notebooks JupyterLab para fins de compatibilidade com versões anteriores, consulte a Instalação das extensões JupyterLab e Jupyter Server.

Repositórios Git

O SageMaker Studio vem pré-instalado com uma extensão Jupyter Git para que os usuários insiram uma URL personalizada de um repositório Git, clonem-na em seu diretório EFS, enviem alterações e visualizem o histórico de confirmações. 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 atualizadas.

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 credenciais ssh em seu diretório EFS individual para facilitar o gerenciamento.

Ambiente Conda

Os notebooks do SageMaker Studio usam o 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 pelo EFS e são persistentes entre as reinicializações do kernel, do aplicativo ou do Studio. O Studio seleciona automaticamente todos os ambientes válidos como kernels KernelGateway.

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 volume Studio EFS em Quatro abordagens para gerenciar pacotes Python em notebooks do Amazon SageMaker Studio.

Conclusão

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

Apêndice

Comparação: multilocação

Tabela 2 — Comparação de multilocação

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 Studio marca automaticamente todos os recursos com o ARN do domínio e o ARN do perfil/espaço do usuário.	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 para o ABAC.
As APIs de lista não podem ser restringidas por tags. A filtragem de recursos da interface do usuário é feita em espaços compartilhados, no entanto, as chamadas da API List feitas por meio do AWS CLI SDK do Boto3 listarão os recursos em toda a região.	O isolamento de APIs de lista também é possível, já que os inquilinos estão em suas contas dedicadas.	As APIs de lista não podem ser restringidas por tags. Listar chamadas de API feitas por meio do AWS CLI SDK do Boto3 listará recursos em toda a região.
SageMaker Os custos de computação e armazenamento do Studio por locatário podem ser facilmente monitorados usando o ARN do domínio como uma etiqueta de alocação de custos.	SageMaker Os custos de computação e armazenamento do Studio por locatário são fáceis de monitorar com uma conta dedicada.	SageMaker Os custos de computação do Studio por inquilino precisam ser calculados usando tags personalizadas. SageMaker Os custos de armazenamento do Studio não podem ser monitorad

Vários domínios	Conta múltipla	Controle de acesso baseado em atributos (ABAC) em um único domínio
		os por domínio, pois todos os locatários compartilham o mesmo volume de EFS.
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 Organizat ions e a venda de várias contas.	O escalonamento precisa de uma função específic a 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 colaboraç ão.

SageMaker Backup e recuperação de domínios do Studio

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

Opção 1: fazer backup do EFS existente usando o 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 no EFS.

   a. Para cada usuário na lista de usuários/espaços, descreva o perfil/espaço do usuário (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 funções de execução distintas.
   d. Identifique a função padrão de execução do Space.
3. Crie um novo domínio e especifique a função 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 de funções de execução.
5. Crie um mapeamento para os novos EFS e UIDs.
   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.

EFS backup

Para fazer backup do EFS, use as instruções a seguir:

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

iii. Alterar a propriedade de todos os arquivos: `chown <new_UID>` de todos os arquivos.

Opção 2: fazer backup do EFS existente 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 funções de execução.
4. No domínio atual do SageMaker Studio, defina um script LCC padrão no nível do domínio.
   - Na LCC, copie tudo para `/home/sagemaker-user` 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 que a LCC seja executada).
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 funções de execução.
9. Configure uma LCC no nível do domínio:
   - Na LCC, copie tudo no prefixo do perfil de usuário no S3 para `/home/sagemaker-user`
10. Crie aplicativos padrão do Jupyter Server para todos os usuários com a configuração da LCC (CLI, SDK).

SageMaker Acesso ao estúdio usando a declaração SAML

Configuração da solução:

1. Crie um aplicativo SAML em seu IdP externo.
2. Configure o IdP externo como um provedor de identidade no IAM.
3. Crie uma função `SAMLValidator` Lambda que possa ser acessada pelo IdP (por meio de uma URL de função ou API Gateway).
4. Crie uma função `GeneratePresignedUrl` Lambda e um API Gateway para acessar a função.
5. Crie uma função do IAM que os usuários possam assumir para invocar o API Gateway. Essa função deve ser passada na declaração SAML como um atributo no seguinte formato:

- Attribute name: https://aws.amazon.com/SAML/Attributes/Role
- Valores de atributo <IdentityProviderARN>, <RoleARN>

6. Atualize o endpoint do SAML Assertion Consumer Service (ACS) para o URL de chamada. SAMLValidator

Código de exemplo do validador SAML:

```
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

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

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 produto do SageMaker, 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.

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.