Como desbloquear o valor dos seus dados em grande escala no setor de serviços financeiros

: Como desbloquear o valor dos seus dados em grande escala no setor de serviços financeiros

Table of Contents

Como desbloquear o valor dos seus dados em grande escala no setor de serviços financeiros

Brian Cavanagh, Maira Ladeira Tanke, Amine Ait el harraj, Junaid Baba, Maren Suilmann, Pauline Ting e Sokratis Kartakis, Amazon Web Services (AWS)

Setembro de 2022 (histórico do documento)

O setor de serviços financeiros está enfrentando uma grande disrupção causada por empresas de tecnologia financeira (fintechs) e bancos somente digitais que estão liderando a transformação digital do setor bancário. Essa transformação é cada vez mais caracterizada pelo desenvolvimento de produtos e serviços bancários baseados em tecnologias de inteligência artificial e aprendizado de máquina (AI/ML). De acordo com AI-bank of the future: Os bancos podem enfrentar o desafio da IA? da McKinsey & Company, as organizações de FS devem implantar AI/ML tecnologias em grande escala se quiserem permanecer relevantes. Para implantar AI/ML tecnologias em grande escala, as organizações de FS devem primeiro descobrir o valor comercial de seus dados.

As organizações de serviços financeiros armazenam grandes quantidades de dados, mas enfrentam dificuldade para extrair valor desses dados em grande escala. Ao revelar o valor dos dados, as organizações de serviços financeiros podem disponibilizar ofertas, insights e suporte mais detalhados e personalizados a seus clientes e parceiros. O valor desbloqueado também pode ajudar as organizações de serviços financeiros a expor rapidamente as ineficiências nos processos atuais, do mercado de capitais às operações de manufatura, ao mesmo tempo que fornece informações priorizadas sobre áreas que exigem otimização. Essa estratégia descreve como você pode liberar o valor dos seus dados em grande escala desenvolvendo recursos de ML em sua organização. O público-alvo dessa estratégia inclui CEOs, CFOs, CIOs e gerentes seniores do setor bancário e de gestão patrimonial.

Essa estratégia ajuda você a entender:

- Resultados comerciais para o lançamento de recursos de ML em sua organização
- Métricas e pontuações desejadas para o sucesso operacional
- Uma estrutura de ML escalável para transformar os recursos de ML da sua organização
- AWS melhores práticas de escalabilidade (com base em centenas de implementações de clientes)

Resultados de negócios desejados

O principal resultado comercial é adotar processos otimizados capazes de transformar os dados da sua organização em valor comercial. Especificamente, essa estratégia pode ajudar você a alcançar os seguintes resultados comerciais específicos:

- Escalar os recursos de operações e machine learning (MLOps) em toda a sua organização e desenvolver uma plataforma de ML para apoiar centenas de cientistas e engenheiros de dados na execução de fluxos de trabalho de ML em sua organização.
- Implemente uma implantação de infraestrutura escalável, segura, econômica e sustentável usando AWS Service Catalog para criar uma infraestrutura gerenciada sob demanda com a Amazon AI. SageMaker
- Padronizar o processo de desenvolvimento e implantação do modelo de ML em várias equipes.
- Reduzir a dívida técnica dos modelos existentes e criar artefatos reutilizáveis para acelerar o desenvolvimento futuro de modelos.
- Reduzir o tempo entre a ideia e a geração de valor dos casos de uso de dados e análises para 12 a 16 semanas.
- Reduzir o tempo de criação de ambientes de casos de uso de ML para 1 a 2 dias.

Quando você escala o uso de análises avançadas na empresa, pode demorar muito tempo para desenvolver e lançar modelos e soluções de ML para produção. Ao fazer parceria com AWS, você recebe orientação e suporte que podem ajudá-lo a projetar, criar e lançar rapidamente uma plataforma de autoatendimento moderna, segura, escalável e sustentável para desenvolver e produzir ML-based serviços para apoiar seus negócios e clientes. AWS Os serviços profissionais podem trabalhar com você para acelerar a implementação das melhores práticas da AWS para usar a SageMaker IA para criar, treinar e implantar modelos de ML para casos de uso personalizados de acordo com suas necessidades organizacionais.

A colaboração possibilita:

- Uma DevOps-driven abordagem federada, de autoatendimento e código de aplicativos, com um caminho claro para a vida útil que pode reduzir o tempo de implantação de semanas para minutos
- Um ambiente seguro, controlado e baseado em modelos para acelerar a inovação com modelos e insights de ML que usam as práticas recomendadas do setor e artefatos compartilhados em todo o banco

- A capacidade de acessar e compartilhar dados com mais facilidade e consistência em toda a organização
- Um conjunto de ferramentas moderno baseado em uma arquitetura gerenciada sob demanda que pode minimizar os requisitos de computação, reduzir os custos e permitir o desenvolvimento e as operações sustentáveis de ML, com a flexibilidade de acomodar novos AWS produtos e serviços para atender aos requisitos contínuos de casos de uso e conformidade
- Suporte a adoção, engajamento e treinamento que permite que as equipes de ciência de dados e engenharia sejam autossuficientes e escalem os produtos de ML em toda a organização

Métricas operacionais

AWS Os Serviços Profissionais podem trabalhar com suas equipes internas para estabelecer um mecanismo de governança de entrega para monitorar o progresso da sua organização em relação às métricas operacionais definidas na tabela a seguir. As métricas são baseadas em valores do mundo real e podem ajudar a definir seu engajamento de MLOps. Os valores na tabela a seguir são apenas diretrizes. Os valores-alvo dependerão da implementação da sua organização.

Área	Subárea	Métrica	Meta de 6 meses a partir do lançamento da plataforma
1. Eficiência operacional	Entrega mais rápida	Tempo para geração de valor	Média de 3 meses
	Transparência de custos por meio de alocações claras de custos para as divisões de negócios	Custos de gerenciamento da AWS	Menos do que os iniciais
2. Boundary-less entrega	Liberdade de acesso concedida de forma simples e controlada	Tempo para acesso	1 dia
	Support for CI/CD	Tempo para geração de valor (do conceito à produção)	Média de 3 meses
	Entrega simplificada do desenvolvimento à produção	Tempo para implantação (após a versão beta)	Média de 3 meses
	On-demand rotação de laboratório up/down	Tempo para inicialização do ambiente	24 horas

3. Acesso	controlado alinhado a cada caso de uso e concedido ou removido dinamicamente	Tempo para acesso a caso de uso	Inferior a 1 dia
	Ambiente de produção persistente que possibilite o desenvolvimento orientado por spokes	Tempo para implantação	2 semanas
	Custos de armazenamento e computação gerenciados no ponto de uso e atribuídos ao centro de custos	Actuals/forecast transparente	Variância permitida da precisão da previsão
	Conformidade com padrões internos de segurança (avaliação e auditoria automáticas)	Número de ambientes fora de conformidade	0
	Proprietários de dados responsáveis por contribuir e compartilhar dados (de acordo com o SLA)	Cópias de dados no data lake	0 cópia de dados no data lake
	Processamento comum com suporte à separação regulatória de dados	Acesso em conformidade	0 descoberta regulatória

4. Padrões repetíveis em constante evolução	Um ambiente de desenvolvimento de modelos com padrões em constante evolução	Número de atualizações por ano	6
	Capacidade de empacotar e implantar ambientes à vontade usando spokes	Tempo para inicialização do ambiente	2 horas
	Layout claro dos serviços e da infraestrutura que estão sendo consumidos	Non-tracked infraestrutura	0
5. Para todo o banco	Número reduzido de ambientes de modelagem distintos	Número de ambientes	< 2
	Colaboração em equipe simplificada ao longo de todos os limites organizacionais	Tempo para acesso (por exemplo, dados)	1 dia
	Operação comum e singular model/framework	Número de domínios controlados pelo modelo	> 3

Visão geral da solução

Uma estrutura de ML escalável

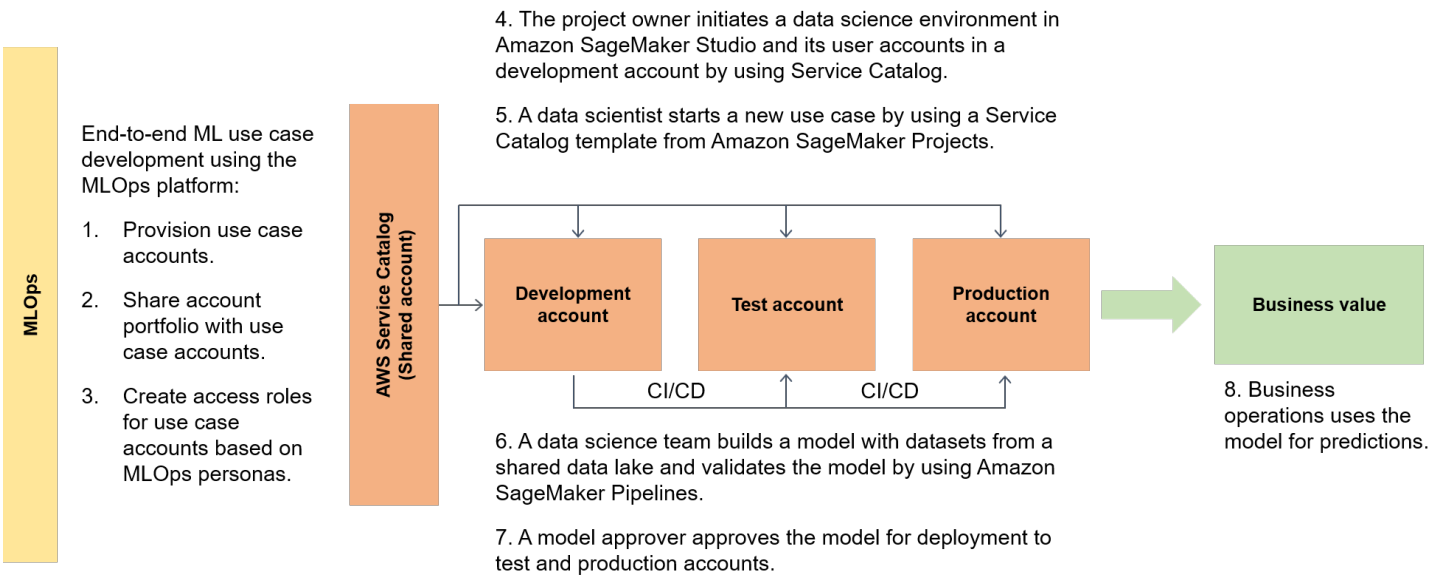
Em uma empresa com milhões de clientes espalhados por várias linhas de negócios, os fluxos de trabalho de ML exigem a integração de dados pertencentes e gerenciados por equipes isoladas usando ferramentas diferentes para gerar valor comercial. Os bancos estão comprometidos com a proteção dos dados de seus clientes. Da mesma forma, a infraestrutura usada para o desenvolvimento de modelos de ML também está sujeita a altos padrões de segurança. Essa segurança adicional aumenta a complexidade e afeta o tempo de valorização dos novos modelos de ML. Em uma estrutura de ML escalável, é possível usar um conjunto de ferramentas modernizado e padronizado para reduzir o esforço necessário para combinar diferentes ferramentas e simplificar o processo do desenvolvimento à produção de novos modelos de ML.

Tradicionalmente, o gerenciamento e o suporte das atividades de ciência de dados no setor de serviços financeiros são controlados por uma equipe de plataforma central que reúne requisitos, provisiona recursos e mantém a infraestrutura para equipes de dados em toda a organização. Para escalar rapidamente o uso de ML em equipes federadas em toda a organização, é possível usar uma estrutura de ML escalável para fornecer recursos de autoatendimento para desenvolvedores de novos modelos e pipelines. Isso permite que esses desenvolvedores implantem uma infraestrutura moderna, pré-aprovada, padronizada e segura. Em última análise, esses recursos de autoatendimento reduzem a dependência da sua organização de equipes de plataformas centralizadas e aceleram a geração de valor para o desenvolvimento de modelos de ML.

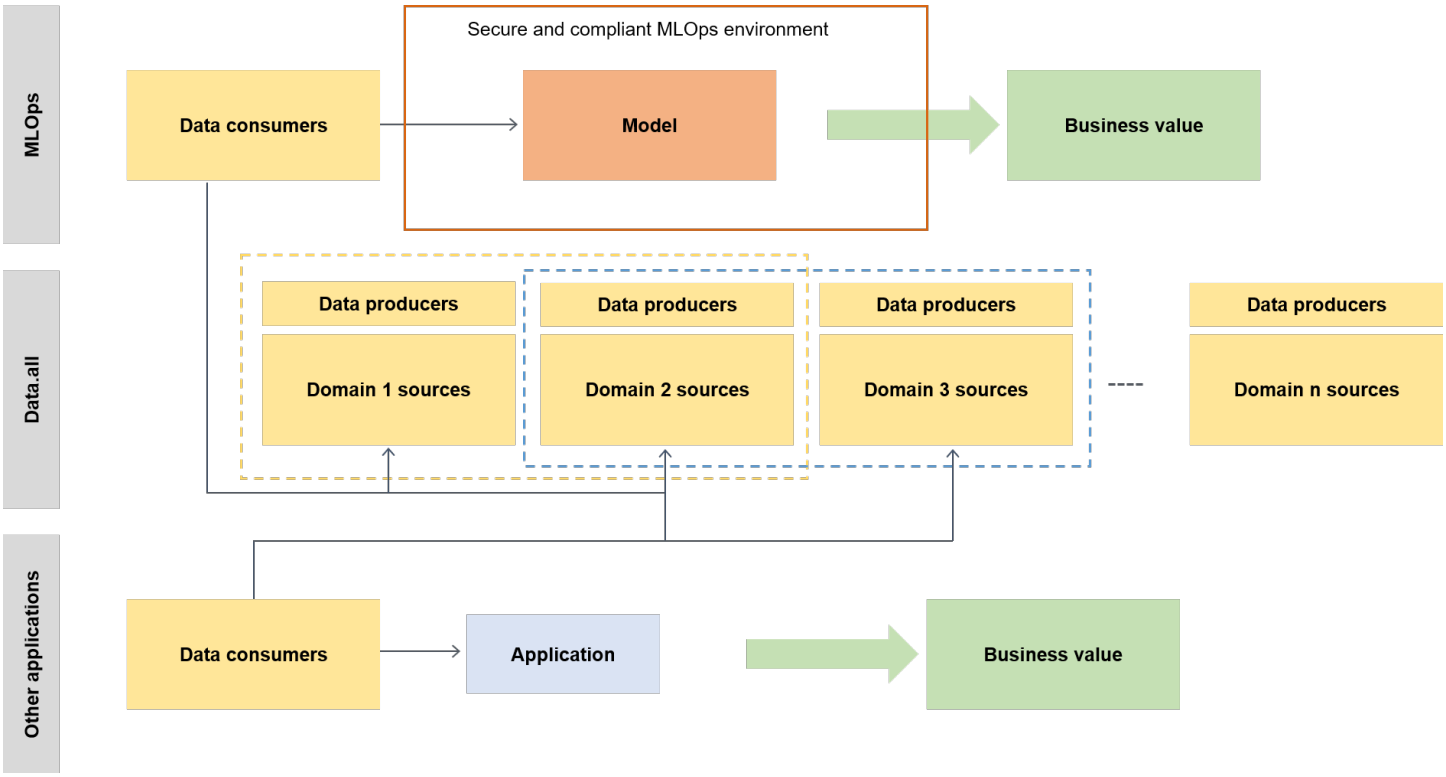
A estrutura de ML escalável permite que os consumidores de dados (por exemplo, cientistas de dados ou engenheiros de ML) liberem valor comercial, o que oferece a eles a capacidade de fazer o seguinte:

- Procurar e descobrir dados pré-aprovados que são necessários para o treinamento de modelos
- Obter acesso a dados pré-aprovados de forma rápida e fácil
- Usar dados pré-aprovados para provar a viabilidade do modelo
- Lançar o modelo comprovado para produção para utilização por outras pessoas

O diagrama a seguir destaca o fluxo ponta a ponta da estrutura e a rota simplificada para os casos de uso de ML.



Em um contexto mais amplo, os consumidores de dados usam um acelerador com tecnologia sem servidor chamado data.all para obter dados em vários data lakes e, em seguida, usam os dados para treinar seus modelos, conforme ilustrado no diagrama a seguir.



Em um nível inferior, a estrutura escalável de ML contém o seguinte:

- Self-service implantação de infraestrutura — reduza sua dependência de equipes centralizadas.

- Sistema central de gerenciamento de pacotes Python: disponibilize pacotes Python pré-aprovados para desenvolvimento de modelos.
- CI/CD pipelines para desenvolvimento e promoção de modelos — Reduza o tempo de vida incluindo integração contínua e pipelines contínuos (CI/CD) como parte de sua infraestrutura como modelos de código (IaC).
- Capacidades de teste de modelos: aproveite as funcionalidades de teste de unidades, teste de modelos, teste de integração e teste ponta a ponta que estão disponíveis automaticamente para novos modelos.
- Desacoplamento e orquestração de modelos — Evite computação desnecessária e torne suas implantações mais robustas desacoplando as etapas do modelo de acordo com os requisitos de recursos computacionais e orquestrando as diferentes etapas usando o Amazon AI Pipelines. SageMaker
- Padronização de código — Melhore a qualidade do seu código usando a integração de CI/CD pipeline para validar os padrões do Python Enhancement Proposal (PEP 8).
- Quick-start modelos genéricos de ML — Obtenha modelos do Service Catalog que instanciam seus ambientes de modelagem de ML (desenvolvimento, pré-produção e produção) e pipelines associados com o clique de um botão usando SageMaker projetos de IA para implantação.
- Monitoramento da qualidade de dados e modelos — Certifique-se de que seus modelos atendam aos requisitos operacionais e dentro do seu nível de tolerância ao risco usando o Amazon SageMaker AI Model Monitor para monitorar automaticamente a variação na qualidade de seus dados e modelos.
- Monitoramento de tendências: permita que os proprietários do seu modelo tomem decisões justas e equitativas verificando automaticamente os desequilíbrios de dados e se as mudanças no mundo introduziram algum viés em seu modelo.

Um hub central para metadados

Data.allé um acelerador sem servidor que você pode integrar aos seus data lakes existentes da AWS para reunir metadados em um hub central. Uma interface de usuário simples e fácil de usar no data.all exibe metadados associados a conjuntos de dados de vários data lakes existentes. Isso permite que usuários não técnicos e técnicos pesquisem, naveguem e solicitem acesso a dados valiosos que podem ser usados em seus laboratórios de ML. Data.all usa AWS Lake Formation, AWS Lambda, Amazon Elastic Container Service (Amazon ECS) AWS Fargate, OpenSearch Amazon Service e. AWS Glue

SageMaker validação

Para provar as capacidades da SageMaker IA em uma variedade de arquiteturas de processamento de dados e ML, a equipe que implementa os recursos seleciona, junto com a equipe de liderança bancária, casos de uso de complexidade variável de diferentes divisões de clientes bancários. Os dados do caso de uso são ofuscados e disponibilizados em um bucket de dados local do Amazon Simple Storage Service (Amazon S3) na conta de desenvolvimento do caso de uso para a fase de comprovação das capacidades.

Quando a migração do modelo do ambiente de treinamento original para uma arquitetura de SageMaker IA estiver concluída, seu data lake hospedado na nuvem disponibilizará os dados para serem lidos pelos modelos de produção. As previsões geradas pelos modelos de produção são então gravadas de volta no data lake.

Depois que os casos de uso candidatos forem migrados, a estrutura escalável de ML usará uma linha de base inicial para as métricas de destino. É possível comparar a linha de base com os horários anteriores de provedores on-premises ou de outros provedores de nuvem como evidência das melhorias de tempo possibilitadas pela estrutura escalável de ML.

Práticas recomendadas

Esta seção fornece uma visão geral das AWS melhores práticas para MLOPs.

Gerenciamento e separação de contas

AWS as melhores práticas para gerenciamento de contas recomendam que você divida suas contas em quatro contas para cada caso de uso: experimentação, desenvolvimento, teste e produção. Também é prática recomendada ter uma conta de governança para fornecer recursos MLOps compartilhados em toda a organização e uma conta de data lake para fornecer acesso centralizado aos dados. A justificativa para isso é separar completamente os ambientes de desenvolvimento, teste e produção, evitar atrasos causados por limites de serviço atingidos por vários casos de uso, evitar que equipes de ciência de dados compartilhem o mesmo conjunto de contas e fornecer uma visão geral completa dos custos de cada caso de uso. Finalmente, é prática recomendada separar os dados no nível da conta, pois cada caso de uso tem seu próprio conjunto de contas.

Padrões de segurança

Para atender a requisitos de segurança, é prática recomendada desativar o acesso público à Internet e criptografar todos os dados com chaves personalizadas. Em seguida, você pode implantar uma instância segura do Amazon SageMaker AI Studio na conta de desenvolvimento em questão de minutos usando o Service Catalog. Você também pode obter recursos de auditoria e monitoramento de modelos para cada caso de uso usando a SageMaker IA por meio de modelos implantados com projetos de SageMaker IA.

Recursos de casos de uso

Depois que a configuração da conta for concluída, os cientistas de dados da sua organização poderão solicitar um novo modelo de caso de uso usando SageMaker projetos de SageMaker IA no AI Studio. Esse processo implanta a infraestrutura necessária para ter recursos de MLOps na conta de desenvolvimento (com o mínimo de suporte exigido das equipes centrais), como CI/CD pipelines, testes unitários, testes de modelos e monitoramento de modelos.

Cada caso de uso é então desenvolvido (ou refatorado no caso de uma base de código de aplicativo existente) para ser executado em uma arquitetura de IA usando SageMaker recursos de SageMaker IA, como rastreamento de experimentos, explicabilidade do modelo, detecção de viés e monitoramento de qualidade. data/model Você pode adicionar esses recursos a cada pipeline de caso de uso usando etapas de pipelines no SageMaker AI Pipelines.

A jornada de maturidade de MLOps

A jornada de maturidade de MLOps define os recursos necessários de MLOps disponibilizados em uma configuração em toda a empresa para garantir que um fluxo de trabalho modelo de ponta a ponta esteja em vigor. A jornada de maturidade consiste em quatro estágios:

1. Inicial: nesse estágio, você estabelece a conta de experimentação. Você também protege uma nova AWS conta em sua organização, onde pode experimentar o SageMaker Studio e outros novos AWS serviços.
2. Repetível: nesse estágio, você padroniza os repositórios de código e o desenvolvimento da solução de ML. Você também adota uma abordagem de implementação de várias contas e padroniza seus repositórios de código para oferecer suporte a governança do modelo e a auditorias de modelos à medida que aumenta a escala da oferta horizontalmente. É prática recomendada adotar uma abordagem de desenvolvimento de modelos pronta para produção com soluções padrão fornecidas por uma conta de governança. Os dados são armazenados em uma conta do data lake e os casos de uso são desenvolvidos em duas contas. A primeira conta é para experimentação durante o período de exploração de ciência de dados. Nessa conta, os cientistas de dados descobrem modelos para resolver o problema de negócios e experimentam várias possibilidades. A outra conta deve ser usada para o desenvolvimento que ocorre após a identificação do melhor modelo e quando a equipe de ciência de dados está pronta para trabalhar no pipeline de inferência.
3. Confiável: nesse estágio, você introduz o teste, a implantação e a implantação de várias contas. É preciso entender os requisitos do MLOps e introduzir testes automatizados. Implemente as práticas recomendadas de MLOps para garantir que os modelos sejam robustos e seguros. Durante essa fase, apresente duas novas contas de casos de uso: uma conta de teste para testar modelos desenvolvidos em um ambiente que emula o ambiente de produção e uma conta de produção para executar a inferência de modelos durante as operações comerciais. Por fim, use testes, implantação e monitoramento automatizados de modelos em uma configuração de várias contas para garantir que seus modelos atendam ao padrão de alta qualidade e performance definidos.
4. Escalável: nesse estágio, você modela e produz várias soluções de ML. Várias equipes e casos de uso de ML começam a adotar MLOps durante o processo de criação de modelos de ponta a ponta. Para obter escalabilidade nesse estágio, você também aumenta o número de modelos em sua biblioteca de modelos por meio de contribuições de uma base mais ampla de cientistas de dados, reduz o tempo de geração de valor da ideia ao modelo de produção para mais equipes em toda a organização e itera à medida que você escala.

Para obter mais informações sobre o modelo de maturidade do MLOps, consulte o roteiro da fundação MLOps para empresas com Amazon AI no blog SageMaker do Machine Learning. AWS

Uma visão de sucesso

Para liberar com sucesso o valor de seus dados, sua organização deve desenvolver uma estrutura de ML escalável que alcance seus resultados comerciais e, ao mesmo tempo, se alinhe às AWS melhores práticas definidas nessa estratégia. Essa estrutura pode ajudar sua organização a garantir:

- Um processo de autoatendimento para criar uma infraestrutura padronizada em nível de produção para desenvolver análises avançadas, cargas de trabalho baseadas em ciência de dados e uma rota clara de sobrevivência para essas cargas de trabalho DevOps-driven
- Equipes de casos de uso capazes de desfrutar de uma governança simplificada, o que reduz o tempo de geração de valor e, ao mesmo tempo, libera recursos técnicos valiosos
- Uma biblioteca de ciência de dados para acessar e compartilhar modelos de ML reutilizáveis que podem reduzir rapidamente os prazos de entrega de ponta a ponta usando modelos que podem ser refinados e compartilhados com outras pessoas
- Uma biblioteca de referência de dados na nuvem (DataHub) que fornece metadados sobre todos os dados localizados em vários lagos de AWS dados, acelerando não apenas a descoberta de dados, mas o tempo de acesso aos dados necessários para treinar modelos
- Uma DevOps equipe dedicada treinada para gerenciar a plataforma e o kit de ferramentas subjacente, apoiar o desenvolvimento de novos recursos e garantir a conformidade contínua com os principais controles
- Real-life casos de uso fornecidos por meio de contas em nível de produção
- Uma transferência de conhecimento entre a equipe do projeto e as equipes da linha de negócios receptora (equipes spoke) que permite que as equipes de negócios autogerenciem o desenvolvimento contínuo e todo processo do desenvolvimento à produção
- Conclusão das atividades iniciais de engajamento e adoção com um grupo-alvo de equipes de negócios
- Centenas de sessões de AWS treinamento em sala de aula ministradas para apoiar a adoção

Perguntas frequentes

Quando devo criar uma plataforma de MLOps?

Padronize em uma plataforma MLOps ao perceber que seus engenheiros estão gastando mais tempo pesquisando e buscando aprovações para opções de ferramentas do que criando modelos de ML.

Posso integrar outras ferramentas de ML na plataforma de MLOps?

Sim. Você pode integrar ferramentas que não são AWS ferramentas na plataforma. Embora o SageMaker AI Studio esteja no centro da plataforma MLOps, você ainda pode integrar outros produtos ao pacote de serviços do SageMaker AI Studio.

Como minha organização pode simplificar os requisitos de governança para acelerar a inovação?

Como parte dos candidatos a casos de uso que você seleciona para provar a construção de sua plataforma de MLOps, certifique-se de que os casos de uso sejam suficientemente complexos, exijam várias classificações de dados e exijam grandes volumes de dados. Ao fazer isso, você não apenas prova a capacidade da plataforma, mas também faz o trabalho pesado do ponto de vista da governança como parte do lançamento inicial da plataforma. Se você puder fazer isso, as equipes que adotarem a plataforma de MLOps como parte da implantação terão uma carga de governança mais leve, pois usam uma plataforma que já atendeu aos requisitos de governança para casos de uso complexos.

De que tipo de equipe eu preciso para criar uma plataforma de MLOps?

Uma base robusta de MLOps, que define claramente a interação entre várias pessoas e tecnologias, pode reduzir o tempo para geração de valor, reduzir custos e permitir que os cientistas de dados se concentrem na inovação. Ter a equipe certa pode ser a diferença entre o fracasso e o sucesso no desenvolvimento da plataforma de MLOps. Devido à natureza dos MLOPs, muitas funções precisam ser envolvidas, como cientistas de dados, engenheiros de ML, DevOps profissionais, proprietários de dados, proprietários de TI, analistas de negócios e proprietários de produtos. Certifique-se de que todas as partes interessadas estejam interagindo em uma equipe multifuncional para garantir os melhores resultados para sua plataforma MLOps.

Como posso começar minha jornada de MLOps?

Comece criando um ambiente de experimentação seguro em que os cientistas de dados recebam um snapshot dos dados. Os cientistas de dados podem usar a SageMaker IA para experimentar e, finalmente, provar que o ML pode resolver um problema comercial específico.

Uma transformação de MLOps deve ser conduzida por uma abordagem de cima para baixo ou de baixo para cima em uma organização?

Embora as abordagens de baixo para cima possam ser bem-sucedidas, o apoio da liderança é essencial para o sucesso do desenvolvimento da plataforma de MLOps. Com uma abordagem de cima para baixo, é possível garantir uma padronização mais rápida da solução desenvolvida, reduzir custos e obter maior escalabilidade e reutilização entre os modelos desenvolvidos por diferentes equipes em sua organização.

Próximas etapas

Para liberar o valor comercial dos seus dados adotando uma estrutura de ML escalável, considere as próximas etapas a seguir:

1. Entre em contato com um representante AWS de vendas ou entre em contato com um representante de serviços AWS profissionais.
2. Busque oportunidades de treinamento e certificação para os membros adequados da equipe em sua organização:
   - AWS treinamento e certificações
   - MLOps Engineering em AWS
   - AWS Big Data certificado - Especialidade
   - AWS Machine Learning certificado — Especialidade
   - AWS DevOps Engenheiro certificado - Profissional
   - AWS Treinamento privado - Solicitação de informações

Recursos

- O que há de novo no Machine Learning? (AWS documentação)
- Automatize MLOPs com projetos de SageMaker IA () YouTube
- Coloque seus dados para trabalhar na nuvem mais escalável, confiável e segura (AWS documentação)
- Roteiro da fundação MLOps para empresas com Amazon SageMaker AI (AWS Machine Learning Blog)
- Parte 1: Como o NatWest Grupo criou uma plataforma MLOps escalável, segura e sustentável (Blog de AWS Machine Learning)
- AI-bank of the future: os bancos podem enfrentar o desafio da IA? (McKinsey e Companhia)

Histórico do documento

A tabela a seguir descreve alterações significativas feitas neste guia. Se desejar receber notificações sobre futuras atualizações, inscreva-se em um feed RSS.

Alteração	Descrição	Data
Substituído DataHub por data.all.	—	7 de novembro de 2023
Publicação inicial	—	16 de setembro de 2022

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.