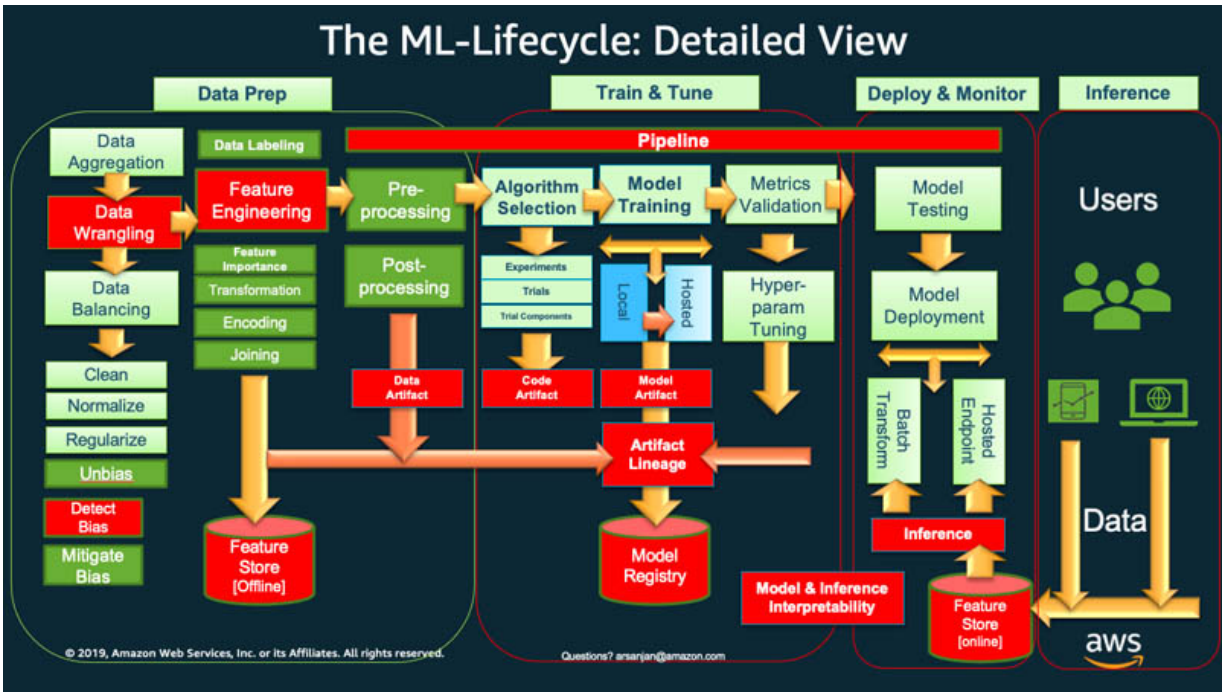
O que é MLOps?

Operações de machine learning (MLOps) são um conjunto de práticas que automatizam e simplificam os fluxos de trabalho e as implantações de machine learning (ML). O machine learning e a inteligência artificial (IA) são recursos essenciais que você pode implementar para resolver problemas complexos do mundo real e agregar valor aos seus clientes. O MLOps é uma cultura e prática de ML que unifica o desenvolvimento de aplicações de ML (Dev) com a implantação e operações do sistema de ML (Ops). Sua organização pode usar o MLOps para automatizar e padronizar processos em todo o ciclo de vida de ML. Esses processos incluem desenvolvimento de modelos, testes, integração, lançamento e gerenciamento de infraestrutura.

Por que o MLOps é necessário?



Em alto nível, para iniciar o ciclo de vida de machine learning, sua organização normalmente precisa começar com a preparação dos dados. Você busca dados de diferentes tipos em várias fontes e realiza atividades como agregação, limpeza de duplicatas e engenharia de recursos.

Depois disso, usa os dados para treinar e validar o modelo de ML. Em seguida, você pode implantar o modelo treinado e validado como um serviço de previsão que outras aplicações podem acessar por meio de APIs.

A análise exploratória de dados geralmente exige que você experimente modelos diferentes até que a melhor versão do modelo esteja pronta para implantação. Isso leva a implantações frequentes de versões de modelos e versões de dados. O rastreamento de experimentos e o gerenciamento do pipeline de treinamento de ML são essenciais antes que suas aplicações possam integrar ou consumir o modelo no seu código.

O MLOps é fundamental para gerenciar de maneira sistemática e simultânea o lançamento de novos modelos de ML com alterações no código e nos dados da aplicação. Uma implementação ideal de MLOps trata os ativos de ML de maneira semelhante a outros ativos de software de ambientes de integração e entrega contínuas (CI/CD). Você implanta modelos de ML junto com as aplicações e os serviços que eles usam e aqueles que os consomem como parte de um processo de lançamento unificado.

Quais são os princípios do MLOps?

A seguir, explicamos quatro princípios fundamentais do MLOps.

Controle de versões

Esse processo envolve o rastreamento de alterações nos ativos de machine learning, para que você possa reproduzir os resultados e reverter para as versões anteriores, se necessário. Cada código de treinamento ou especificação de modelo de ML passa por uma fase de revisão de código. Cada um é versionado para tornar o treinamento de modelos de ML reproduzível e auditável.

A reprodutibilidade em um fluxo de trabalho de ML é importante em todas as fases, desde o processamento de dados até a implantação do modelo de ML. Isso significa que cada fase deve produzir resultados idênticos com as mesmas informações.

Automação

Automatize vários estágios no pipeline de machine learning para garantir repetibilidade, consistência e escalabilidade. Isso inclui estágios desde a ingestão de dados, o pré-processamento, o treinamento de modelos e a validação até a implantação.

Esses são alguns fatores que podem acionar o treinamento e a implantação automatizados de modelos:

- Sistema de mensagens
- Monitoramento ou eventos de calendário
- Alterações de dados
- Alterações no código de treinamento de modelos
- Alterações no código da aplicação.

Testes automatizados ajudam você a descobrir problemas com antecedência para corrigir erros e aprender rapidamente. A automação é mais eficiente com a infraestrutura como código (IaC). Você pode usar ferramentas para definir e gerenciar a infraestrutura. Isso ajuda a garantir que seja ela reproduzível e possa ser implantado de forma consistente em vários ambientes.

Leia sobre a IaC »

X contínuo

Por meio da automação, você pode executar testes continuamente e implantar código no seu pipeline de ML.

Em MLOps, *contínuo* refere-se a quatro atividades que acontecem continuamente quando alguma alteração é feita em qualquer lugar do sistema:

- A *integração contínua* estende a validação e os testes do código aos dados e modelos no pipeline

- A *entrega contínua* implanta automaticamente o modelo recém-treinado ou o serviço de previsão de modelos
- O *treinamento contínuo* retreina automaticamente os modelos de ML para reimplantação
- O *monitoramento contínuo* diz respeito ao monitoramento de dados e ao monitoramento de modelos usando métricas relacionadas aos negócios

Governança de modelos

Governança envolve o gerenciamento de todos os aspectos dos sistemas de ML para fins de eficiência. Você deve realizar muitas atividades para governança:

- Promover a estreita colaboração entre cientistas de dados, engenheiros e partes interessadas dos negócios
- Usar documentação clara e canais de comunicação eficientes para garantir que todos estejam alinhados
- Estabelecer mecanismos para coletar feedback sobre as previsões do modelo e treinar ainda mais os modelos
- Garantir que dados confidenciais estejam protegidos, o acesso aos modelos e à infraestrutura seja seguro e os requisitos de conformidade sejam atendidos

Também é essencial ter um processo estruturado para revisar, validar e aprovar os modelos antes que estes sejam lançados. Isso pode envolver a verificação de justiça, parcialidade e considerações éticas.

Quais são os benefícios do MLOps?

O machine learning ajuda as organizações a analisar dados e obter insights para a tomada de decisões. No entanto, trata-se de um campo inovador e experimental que vem com seu próprio conjunto de desafios. Proteção de dados confidenciais, pequenos orçamentos, escassez de habilidades e tecnologias em constante evolução limitam o sucesso de um projeto. Sem controle e orientação, os custos podem aumentar e as equipes de ciência de dados podem não alcançar os resultados desejados.

O MLOps fornece um mapa para orientar os projetos de ML em direção ao sucesso, independentemente das restrições. Estes são alguns dos principais benefícios dos MLOps.

Menor tempo de introdução no mercado

O MLOps fornece à sua organização uma framework para atingir suas metas de ciência de dados com mais rapidez e eficiência. Seus desenvolvedores e gerentes podem se tornar mais estratégicos e ágeis no gerenciamento de modelos. Os engenheiros de ML podem provisionar a infraestrutura por meio de arquivos de configuração declarativos para iniciar projetos com mais facilidade.

A automação da criação e implantação de modelos resulta em tempos mais rápidos para entrada no mercado com

menos custos operacionais. Os cientistas de dados podem explorar rapidamente os dados de uma organização para oferecer mais valor comercial a todos.

Produtividade aprimorada

As práticas de MLOps aumentam a produtividade e aceleram o desenvolvimento de modelos de ML. Por exemplo, você pode padronizar o ambiente de desenvolvimento ou experimento. Em seguida, seus engenheiros de ML podem lançar novos projetos, alternar entre projetos e reutilizar modelos de ML em todas as aplicações. Eles podem criar processos repetíveis para experimentação rápida e treinamento de modelos. As equipes de engenharia de software podem colaborar e coordenar o ciclo de vida de desenvolvimento de softwares de ML para obter maior eficiência.

Implantação eficiente de modelos

O MLOps melhora a solução de problemas e o gerenciamento de modelos em produção. Por exemplo, engenheiros de software podem monitorar a performance de modelos e reproduzir o comportamento para solucionar problemas. Eles podem rastrear e gerenciar centralmente as versões dos modelos e escolher a correta para diferentes casos de uso comercial.

Ao integrar fluxos de trabalho de modelos com pipelines de integração contínua e entrega contínua (CI/CD), você limita a

degradação da performance e mantém a qualidade do seu modelo. Isso é válido mesmo após upgrades e ajustes de modelos.

Como implementar o MLOps na organização

Há três níveis de implementação de MLOps, dependendo da maturidade da automação na sua organização.

MLOps de nível 0

Fluxos de trabalho manuais de ML e um processo orientado por cientistas de dados caracterizam o *nível 0* para organizações que estão começando com sistemas de machine learning.

Cada etapa é manual, incluindo preparação de dados, treinamento de ML e performance e validação de modelos. Isso exige uma transição manual entre as etapas, e cada etapa é executada e gerenciada de maneira interativa. Os cientistas de dados normalmente entregam modelos treinados como artefatos que a equipe de engenharia implanta na infraestrutura da API.

O processo separa os cientistas de dados que criam o modelo e os engenheiros que o implantam. Lançamentos pouco frequentes significam que as equipes de ciência de dados podem retreinar modelos apenas algumas vezes por ano. Não há considerações de CI/CD para modelos de ML

com o restante do código da aplicação. Da mesma forma, o monitoramento ativo da performance é inexistente.

MLOps de nível 1

As organizações que desejam treinar os mesmos modelos com novos dados frequentemente precisam de uma implementação de maturidade de *nível 1*. O MLOps nível 1 visa treinar o modelo continuamente, automatizando o pipeline de ML.

No nível 0, você implanta um modelo treinado em produção. Diferentemente, para o nível 1, você implanta um pipeline de treinamento que é executado recorrentemente para fornecer o modelo treinado às suas outras aplicações. No mínimo, você obtém a entrega contínua do serviço de previsão de modelos.

A maturidade de nível 1 tem as seguintes características:

- Etapas rápidas do experimento de ML que envolvem automação significativa
- Treinamento contínuo do modelo em produção com dados atualizados como acionadores ativos do pipeline
- Implementação do mesmo pipeline em ambientes de desenvolvimento, pré-produção e produção

Suas equipes de engenharia trabalham com cientistas de dados para criar componentes de código modularizados que são reutilizáveis, combináveis e potencialmente compartilháveis em todos os pipelines de ML. Você também

cria um armazenamento de recursos centralizado que padroniza o armazenamento, o acesso e a definição de recursos para treinamento e atendimento de ML. Além disso, você pode gerenciar metadados, como informações sobre cada execução do pipeline e dados de reprodutibilidade.

MLOps de nível 2

O MLOps *nível 2* é para organizações que desejam experimentar mais e criar com frequência novos modelos que exigem treinamento contínuo. Ele é adequado para empresas orientadas por tecnologia que atualizam seus modelos em minutos, os retreinam a cada hora ou diariamente e os reimplantam simultaneamente em milhares de servidores.

Como há vários pipelines de ML em ação, uma configuração de MLOps de nível 2 requer toda a configuração de MLOps de nível 1. Também requer o seguinte:

- Um orquestrador de pipeline de ML
- Um registro de modelos para rastrear vários modelos

Os três estágios a seguir se repetem em grande escala para vários pipelines de ML a fim de garantir a entrega contínua de modelos.

Construir o pipeline

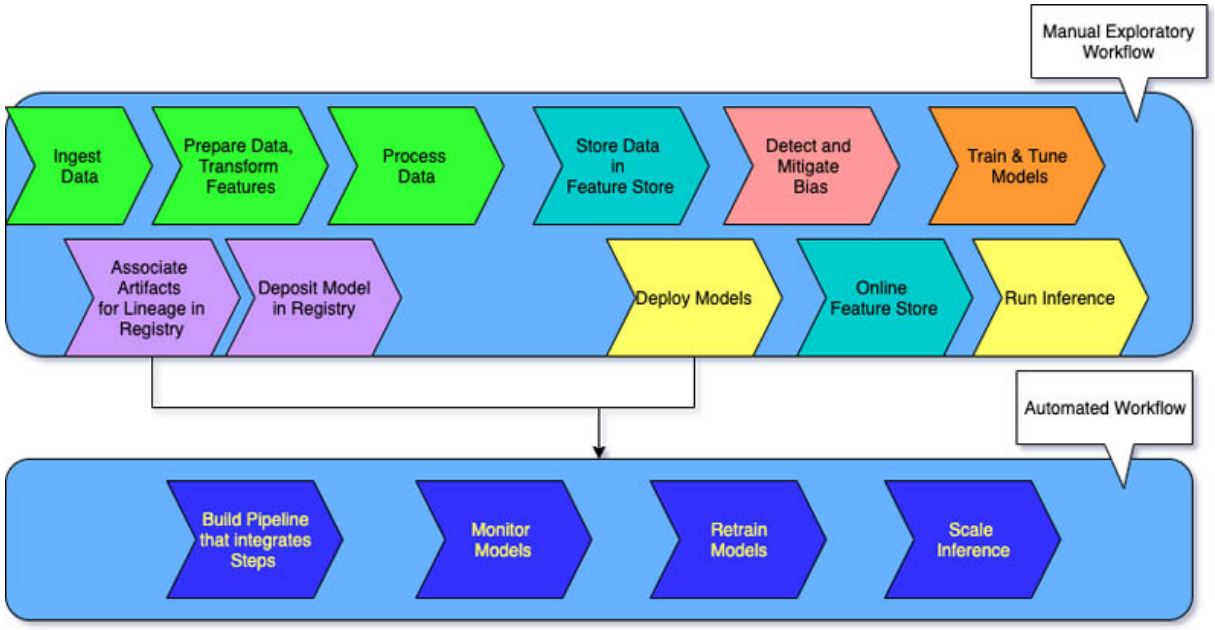
Você experimenta iterativamente novas modelagens e novos algoritmos de ML, garantindo que as etapas do experimento sejam orquestradas. Esse estágio gera o código-fonte para seus pipelines de ML. Você armazena o código em um repositório de origem.

Implantar o pipeline

Em seguida, você cria o código-fonte e executa testes para obter componentes do pipeline para implantação. O resultado é um pipeline implantado com a implementação do novo modelo.

Fornecer o pipeline

Por fim, você fornece o pipeline como um serviço de previsão para suas aplicações. Você coleta estatísticas sobre o serviço de previsão de modelo implantado a partir de dados ativos. Essa saída do estágio é um gatilho para executar o pipeline ou um novo ciclo de experimentos.



Qual é a diferença entre MLOps e DevOps?

MLOps e DevOps são práticas que visam melhorar os processos nos quais você desenvolve, implanta e monitora aplicações de software.

O DevOps visa preencher a lacuna entre as equipes de desenvolvimento e operações. O DevOps ajuda a garantir que as alterações no código sejam automaticamente testadas, integradas e implantadas em produção de maneira eficiente e confiável. Ele promove uma cultura de colaboração para alcançar ciclos de lançamento mais rápidos, melhor qualidade das aplicações e uso mais eficiente dos recursos.

O MLOps, por outro lado, é um conjunto de melhores práticas projetado especificamente para projetos de machine learning. Embora possa ser relativamente simples

implantar e integrar softwares tradicionais, os modelos de ML apresentam desafios únicos. Eles envolvem coleta de dados, treinamento de modelos, validação, implantação e monitoramento e reciclagem contínuos.

O MLOps se concentra em automatizar o ciclo de vida do ML. Isso ajuda a garantir que os modelos não só sejam desenvolvidos, como também implantados, monitorados e retreinados de maneira sistemática e repetida. Ele traz os princípios do DevOps para o ML. O MLOps resulta em uma implantação mais rápida de modelos de ML, maior precisão ao longo do tempo e maior garantia de que eles fornecerão valor comercial real.

Como a AWS pode oferecer suporte aos seus requisitos de MLOps?

O Amazon SageMaker é um serviço totalmente gerenciado que você pode usar para preparar dados e criar, treinar e implantar modelos de ML. Ele é adequado para qualquer caso de uso com infraestrutura, ferramentas e fluxos de trabalho totalmente gerenciados.

O SageMaker fornece ferramentas específicas para MLOps para automatizar processos em todo o ciclo de vida do ML. Ao usar as ferramentas do Sagemaker para MLOps, você pode atingir rapidamente o nível 2 de maturidade de MLOps em grande escala.

Estes são os principais recursos do SageMaker que você pode usar:

- Use o SageMaker Experiments para rastrear artefatos relacionados aos seus trabalhos de treinamento de modelos, como parâmetros, métricas e conjuntos de dados.
- Configure os pipelines do SageMaker para serem executados automaticamente em intervalos regulares ou quando determinados eventos forem acionados.
- Use o SageMaker Model Registry para rastrear versões de modelos. Também é possível rastrear seus metadados, como o agrupamento de casos de uso, e modelar linhas de base de métricas de performance em um repositório central. Você pode usar essas informações para escolher o melhor modelo com base nos seus requisitos de negócios.

Comece a usar o MLOps na Amazon Web Services (AWS) criando uma conta hoje mesmo.