[Aula 4 - Deply - Teórica]

1 Qual é a importância de se automatizar o fluxo de ML?
A importância de se automatizar o fluxo de ML está em garantir que o modelo funcione de forma confiável, previsível, escalável e contínua. Para orientar este processo, existem 3 princípios:
escalabilidade, reprodutibilidade e monitoramento.

2 Elenque casos de uso onde você deve oferecer predições online ou predições em batch. **Necessário citar no minimo dois exemplos de cada.**
Online: previsões são geradas **imediatamente** quando o sistema recebe uma requisição
Exemplos:
Google Tradutor

- Você digita uma frase em inglês.
- Em milissegundos, o modelo retorna a tradução em francês.

Detecção de Fraude em Cartões de Crédito

- Você passa o cartão em uma máquina.
- O modelo deve processar a transação assim que ela ocorre e decidir aprovar ou bloquear a operação em milissegundos.

Chatbot Ifood

- Você digita: *Onde está meu pedido?*
- Em milissegundos, o chatbot processa sua mensagem, consulta seus pedidos e responde com o status atualizado.

Batch: são geradas **periodicamente** ou **sob demanda**, e **armazenadas** para uso posterior. As predições são feitas **antes** da solicitação do usuário.
Exemplos:
Netflix

- O sistema gera recomendações de filmes **para todos os usuários** a cada 4 horas.
- Quando o usuário entra na plataforma, as recomendações **já estão prontas**.

Score de crédito

- Durante a madrugada do primeiro dia de cada mês, o sistema roda um job que recalcula o score de crédito de todos os clientes.
- O job processa milhões de perfis em algumas horas.
- Na manhã seguinte, o novo score já está disponível para uso interno, como concessão de empréstimos ou limites de cartão.

3 Faça um breve comparativo sobre os níveis de maturidade do ciclo de

MLops que são possíveis de serem atingidos nas empresas.

Característica	Nível 0	Nível 1	Nível 2
Estado	Iniciante	Transição	Maturidade
Processos	Manuais	Automatizados	Automatizados
Deploy	Transição manual	Automatizados	Automatizados
Equipe	Separação clara em quem cria e quem implanta	Colaboração, componentes modulares e reutilizáveis	Colaboração, componentes modulares e reutilizáveis
Entregas	Baixa frequência	Alta frequência	Alta frequência
CI/CD	Não existe	Existe	Existe
Reprodutibilidade	Baixa	Alta	Alta
Monitoramento	Não existe	Básico	Avançado
Orquestrador de pipelines	Não existe	Não existe	Existe
Registro de modelos	Não existe	Não existe	Existe

4 Qual é a vantagem de já se ter features pré-processadas em um feature store?
A principal vantagem é permitir a reutilização e consistência entre diferentes modelos de Machine Learning

5 Cite casos de uso de um feature store online e offline.

Online:

- **Previsões em tempo real**: Aplicações como detecção de fraudes, sistemas de recomendação, ou preços dinâmicos exigem acesso instantâneo a recursos para fazer previsões oportunas.
- **Aplicações interativas**: Aplicativos voltados para o usuário que precisam se adaptar rapidamente com base no comportamento do usuário ou em fatores ambientais se beneficiam de lojas de recursos on-line.
- **Teste A/B**: Executar experimentos onde os recursos precisam ser servidos de forma consistente em diferentes segmentos de usuários em tempo real.

Offline:

- **Treinamento de modelo**: Ao desenvolver novos modelos ou retreinar

os existentes, os armazenamentos de recursos offline fornecem os recursos históricos necessários para informar o processo de treinamento.
- **Pontuação em lote**: Em situações em que as previsões são necessárias para um grande conjunto de dados e não em tempo real, os armazenamentos de recursos offline permitem uma pontuação eficiente.
- **Exploração de dados**: Cientistas de dados podem usar armazenamentos offline para análise exploratória de dados, ajudando-os a entender a importância dos recursos antes de passar para a produção.

6 Quais tecnlogias são possíveis de serem utilizadas para se montar um feature store online e offline. Justifique as escolhas de cada recurso. *OBS: Pode usar de referência produtos AWS.

Feature Store Offline

1. Amazon S3
- Armazena datasets históricos, particionados e versionados.
- Baixíssimo custo e alta durabilidade.

2. AWS Glue / EMR / Spark
- Processamento distribuído para criação de features agregadas.
- Bom para jobs pesados e ETL.

3. Athena
- Consulta de features históricas sem necessidade de cluster ativo.

Justificativa:
- Esses serviços permitem processar petabytes de dados com custo baixo e escalabilidade automática, ideais para pipelines de treinamento.

Feature Store Online

1. Amazon DynamoDB
- Latência < 10 ms.
- Ideal para leitura de features para predições online.

2. Amazon ElastiCache (Redis)
- Latência extremamente baixa (< 1 ms).
- Útil quando o modelo precisa de respostas absurdamente rápidas.

3. Amazon API Gateway + Lambda
- Para servir o modelo usando features consultadas no store.

4. Amazon SageMaker Feature Store
- Solução da AWS que combina:
  - store offline (S3)
  - store online (low-latency DB)
  - catalogação
  - versionamento

Justificativa:

- DynamoDB escala automaticamente, é serverless e oferece velocidade.
- Redis atende casos críticos de latência.
- SageMaker Feature Store simplifica toda a arquitetura combinando online e offline.