Pipeline de Ciência de Dados – Spotify

1. Coleta e Ingestão de Dados

O Spotify captura eventos em tempo real:

- Músicas reproduzidas até o fim.
- Músicas puladas.
- Curtidas e adição em playlists.
- Pesquisas por artistas ou faixas.
  Esses dados são enviados para um **Data Lake**, em formato bruto (logs de eventos).

---

2. Armazenamento e Gerenciamento

- **Data Lake** → guarda os dados originais (logs).
- **Data Warehouse** → organiza dados estruturados, como:
    - Perfil do usuário (idade, localização, preferências).
    - Metadados da música (gênero, artista, BPM, popularidade).

---

3. Processamento e Transformação (Engenharia de Atributos)

Transforma dados brutos em atributos relevantes para o modelo:

- **Taxa de Pulo**: quantas vezes o usuário pulou músicas.
- **Hora do Dia**: se escuta mais de manhã, tarde ou noite.
- **Sequência de Gêneros**: padrões de mudança de estilo (ex.: Rock → Jazz).
- **Tempo de Engajamento**: tempo médio de escuta por faixa.

---

4. Análise e Modelagem

- Modelos de **filtragem colaborativa** e **deep learning** para recomendações.
- Treinados com milhões de interações.
- O modelo prevê a probabilidade de um usuário gostar de uma música nova.

---

5. Visualização e Apresentação

- Para o usuário → playlists personalizadas como **“Descobertas da Semana”** ou **“Feito para Você”**.
- Para a empresa → dashboards internos com métricas como:
    - Taxa de aceitação das recomendações.
    - Horas de música ouvidas por semana.
    - Retenção e engajamento do usuário.