

Spotify Data Streaming

Guia de Projeto em Engenharia de Dados

Este documento apresenta o guia para o desenvolvimento do projeto **Spotify Data Streaming**, um projeto de aprendizagem na área de Engenharia de Dados com foco na recolha, processamento e visualização de dados de streaming provenientes da API do Spotify. O objetivo é fornecer uma direção técnica clara para implementar um pipeline de dados funcional, moderno e alinhado com práticas do mercado.

Objetivos do Projeto:

- Desenvolver um pipeline de dados que recolha informações da conta pessoal do Spotify em tempo real.
- Processar e transformar os dados para análise e visualização.
- Construir um dashboard interativo que apresente métricas de streaming, artistas, músicas e horários de reprodução.
- Praticar o uso de ferramentas essenciais de Engenharia de Dados: ingestão, processamento, armazenamento e visualização.
- Compreender conceitos de ETL/ELT e streaming em tempo real.

Arquitetura e Componentes do Sistema

O projeto será composto por quatro camadas principais, simulando uma arquitetura típica de dados moderna:

1. **Ingestão de Dados:** utilização da API do Spotify para obter dados de reprodução e estatísticas do utilizador.
2. **Processamento:** tratamento e transformação de dados usando Python e bibliotecas de manipulação (Pandas, JSON, etc.).
3. **Armazenamento:** persistência dos dados num sistema de base de dados relacional (PostgreSQL) ou em formato parquet/CSV.
4. **Visualização:** criação de dashboards em tempo real com Streamlit ou Grafana.

Tecnologias e Ferramentas

Categoria	Ferramentas / Tecnologias
Linguagem principal	Python
APIs e Bibliotecas	Spotify Web API, Spotipy, Requests
Processamento de dados	Pandas, JSON, PySpark (opcional)
Base de dados	PostgreSQL / SQLite
Orquestração (opcional)	Apache Airflow (local, com Docker)
Visualização	Streamlit / Grafana
Cloud (futuro)	AWS S3, Lambda ou GCP Storage e Functions

Plano de Desenvolvimento

O desenvolvimento do projeto pode ser dividido em fases sequenciais para facilitar o progresso e aprendizagem:

Fase 1 — Configuração Inicial:

- Criar credenciais e acesso à Spotify Web API.
- Configurar ambiente de desenvolvimento (Python, virtualenv, dependências).

Fase 2 — Ingestão de Dados:

- Implementar scripts de extração usando Spotipy.
- Recolher dados de reprodução recentes e histórico pessoal.

Fase 3 — Processamento e Armazenamento:

- Normalizar dados (formatos, timestamps, duplicados).
- Armazenar num banco de dados relacional ou ficheiros parquet.

Fase 4 — Visualização:

- Criar dashboard com Streamlit (músicas mais ouvidas, artistas, horários).
- Implementar atualização automática (simulação de streaming em tempo real).

Fase 5 — Otimização e Automação (opcional):

- Integrar Apache Airflow para orquestração de tarefas ETL.
- Considerar deploy parcial em cloud (ex: AWS Free Tier).

Competências a Desenvolver

- Interação com APIs REST e autenticação OAuth.
- Manipulação e limpeza de dados em Python.
- Criação de pipelines ETL simples e escaláveis.
- Conceitos básicos de dados em tempo real (streaming).
- Criação de dashboards interativos com ferramentas modernas.
- Compreensão de arquiteturas de dados modernas e orquestração.

Conclusão

O projeto **Spotify Data Streaming** é uma base sólida para iniciar uma trajetória em Engenharia de Dados. Permite adquirir experiência prática em ETL, APIs e visualização, com uma abordagem progressiva que pode ser expandida para cloud e pipelines de produção à medida que o conhecimento evolui.