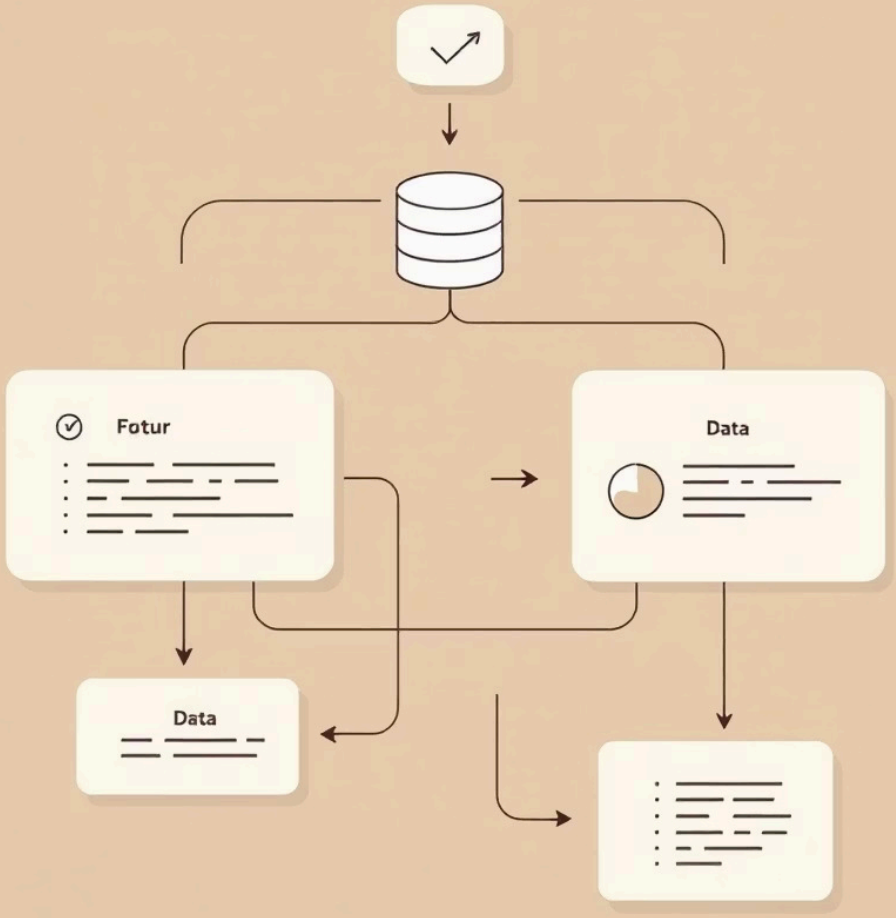
Pipeline de Dados — The Movies Dataset

Ingestão, Armazenamento e Transformação de Dados com Python (Kaggle Dataset)

Relatório Técnico – 2025



Tema da Apresentação

Pipeline de Dados

Estrutura para processamento automatizado.

The Movies Dataset

Conjunto de dados brutos do Kaggle.

RAW → Processed

Transição da camada bruta para a camada limpa.

Esta apresentação detalha a arquitetura e a execução de um pipeline de dados fundamental, focando nas etapas de obtenção, limpeza e integração dos dados do **The Movies Dataset**.

Estrutura do Projeto: Fluxo Arquitetural

Ingestão

Obtenção dos dados brutos do Kaggle (RAW).

Armazenamento

Definição das camadas de dados (Bronze/Silver).

Transformação

Limpeza, integração e enriquecimento dos dados (Processed).

Consumo

Dados prontos para análise e relatórios.

Visão Geral do Projeto

O objetivo principal é construir uma arquitetura de dados robusta e modular para processar o **The Movies Dataset** do Kaggle, garantindo que os dados estejam limpos e prontos para consumo analítico.

> O pipeline foi desenvolvido para obter, limpar e integrar dados do The Movies Dataset (Kaggle). As etapas principais são: Ingestão → Armazenamento → Transformação.

Etapa 1 – Ingestão de Dados (RAW Layer)

Nesta fase, os arquivos de origem são materializados e armazenados em sua forma original, estabelecendo a **Camada Bruta (RAW/Bronze)**.

Origem

Kaggle: **The Movies Dataset**

Script

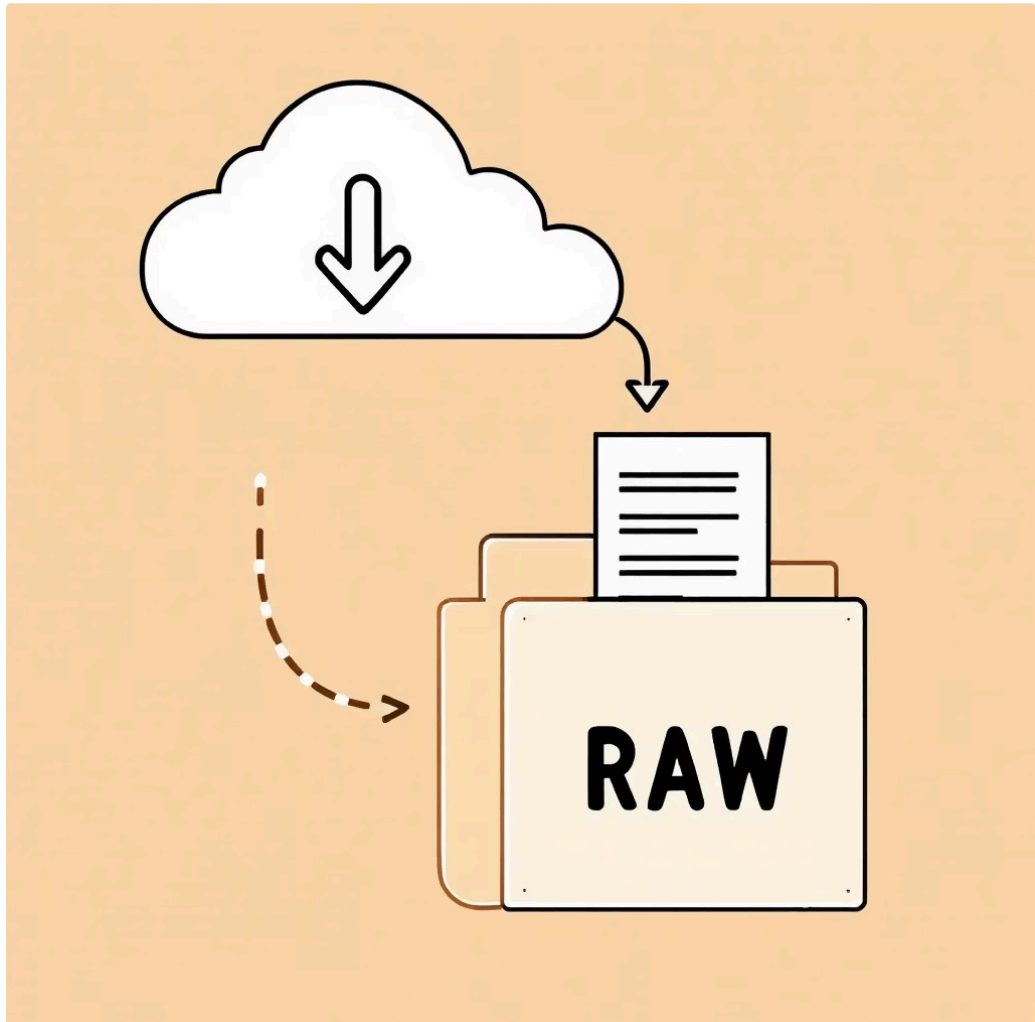
`src/01_ingestion.ipynb`

Saída

`dados/raw/` (Materialização da camada bruta)

Entradas:

- movies_metadata.csv
- credits.csv



Etapa 2 – Armazenamento: Definição das Camadas

O armazenamento é estruturado em duas camadas principais, garantindo rastreabilidade e qualidade dos dados ao longo do pipeline.

RAW (Bronze)

Armazena os dados brutos, inalterados, diretamente da fonte. Fonte única da verdade.

- `dados/raw/`

Processed (Silver)

Contém dados que foram limpos, validados e integrados. Prontos para consumo analítico.

- `dados/processed/`

PS: Futuramente iremos migrar o armazenamento de CSV para o formato **Parquet** para otimizar o desempenho de leitura e compressão.

Etapa 3 – Transformação e Integração

Nesta etapa, a lógica de negócio é aplicada para refinar os dados brutos e prepará-los para análise, gerando o conjunto de dados final na camada **Processed**.

Limpeza e Tipagem

Tratamento de colunas nulas, conversão de tipos e remoção de registros inválidos no dataset **movies**.

Extração

Conversão das estruturas aninhadas nos dados de **credits** para extrair **director** e **main_actor**.

Integração de Dados

Realização do **merge** (junção) dos datasets **movies** e **credits** utilizando o identificador único **id**.

O script `src/02_transformation.ipynb` recebe arquivos da camada RAW e gera a saída `dados/processed/movies_full_cleaned.csv`.

Saída do Pipeline: Dados Prontos para Análise

O resultado final é um conjunto de dados integrado e limpo, armazenado na **Camada Processed (Silver)**, que serve como fonte de dados para dashboards e modelos preditivos.

Destino	Arquivo	Camada
Final	movies_full_cleaned.csv	Processed (Silver)

Colunas Processadas (Exemplo)

- **id, title, runtime, vote_average, vote_count**
- **original_language, year, genre_name**
- director, main_actor

Incrementos Futuramente Planejados

Para aumentar a performance, a usabilidade e o valor analítico do conjunto de dados, estão planejadas as seguintes melhorias:

Formato Parquet

Migração de CSV para Parquet (**to_parquet**) para leitura mais rápida e otimização de armazenamento.

Feature Engineering

Criação de novas variáveis preditivas

GOLD

Resumo do Pipeline: Do Dado Bruto à Informação Confiável