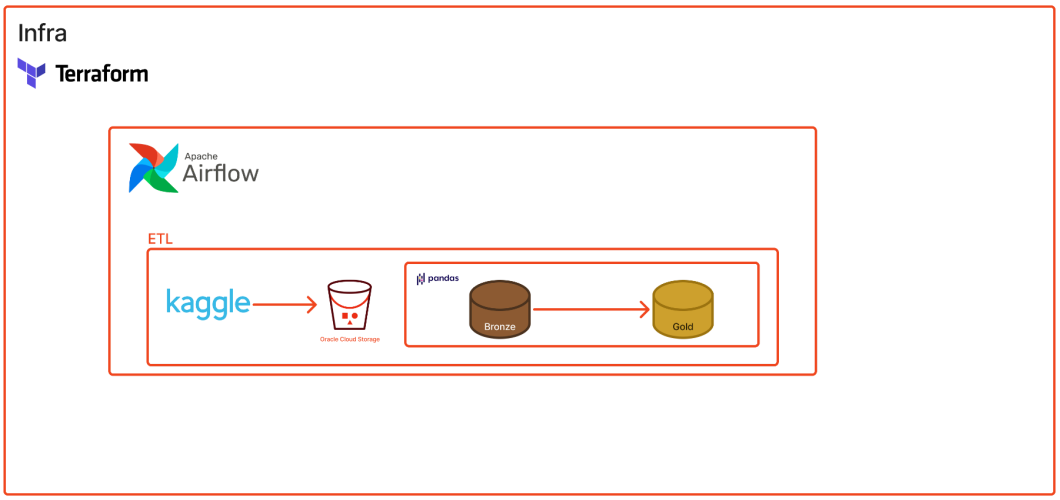
IBM HR Analytics

Este projeto implementa um pipeline ETL orquestrado pelo Apache Airflow, utilizando Python e pandas para processamento de dados, com armazenamento intermediário no Oracle Cloud Object Storage e saída final em CSV para análise no Power BI.



Fluxo do Pipeline

Extração

O script `extract.py` realiza o download do conjunto de dados do Kaggle utilizando a biblioteca `kagglehub`. Os arquivos são descompactados/copiados para o diretório `raw`.

Armazenamento em Nuvem

Os dados podem ser enviados para o Oracle Cloud Object Storage para centralização, versionamento e compartilhamento seguro.

Processamento (ETL)

O Apache Airflow orquestra as tarefas de ETL. O processamento das camadas Bronze → Silver → Gold é executado utilizando a biblioteca pandas. O resultado final é salvo no arquivo `attrition_metrics.csv`.

Bronze: Armazena os dados brutos, sem tratamento, para garantir reprocessamento e auditoria.

Gold: Agrega, calcula métricas (como taxa de rotatividade por departamento) e gera um

dataset enxuto, pronto para consumo analítico.

O uso do pandas permite flexibilidade, rapidez e fácil manutenção do código de transformação.

Análise

O arquivo CSV pode ser importado diretamente no Power BI para visualização e análise.

Como Executar

1. **Clonar o repositório e configurar as variáveis de ambiente:**
   Preencha o arquivo `.env` com as credenciais e configurações pertinentes.
2. **Iniciar os contêineres:**
   Execute o script de extração.
3. **Acessar a interface do Airflow:**
   Disponível em `http://localhost:8080`.
4. **Executar o DAG `transform_gold`:**
   Este procedimento gerará o arquivo CSV final.
5. **Abrir o Power BI:**
   Importe o arquivo `attrition_metrics.csv` para a criação dos painéis de controle.