Projeto de Engenharia de Dados

Pipeline ETL Completo para Análise de Mercado de Trabalho

O que é o Projeto?

Este projeto é um pipeline ETL (Extrair, Transformar, Carregar) que transforma dados brutos em informações limpas e prontas para análise, usando uma arquitetura em camadas.

1

Extrair

Coleta de dados brutos.

2

Transformar

Limpeza e organização dos dados.

3

Carregar

Entrega de dados para análise.

Qual é o Objetivo?

Transformar dados bagunçados em dados úteis, criando um fluxo automático e confiável para análise do mercado de trabalho.

1

Organizar Dados

Estruturar informações do mercado de trabalho.

2

Fluxo Automático

Criar um tratamento de dados eficiente.

3

Boas Práticas

Demonstrar engenharia de dados de qualidade.

4

Tabelas Confiáveis

Produzir dados finais para análises precisas.

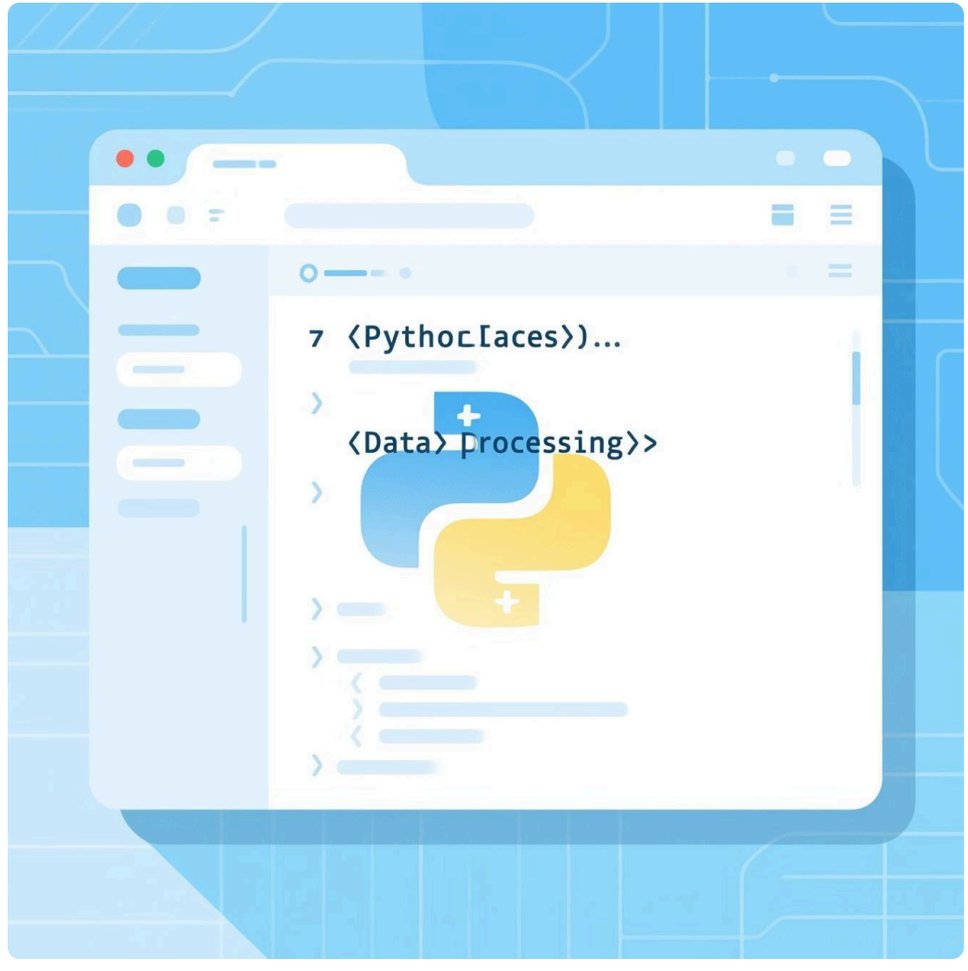
Estrutura do Projeto: Pastas Principais

data/

Contém os arquivos CSV originais (jobs.csv, companies.csv, skills.csv) com informações reais sobre vagas, empresas, salários e habilidades.

scripts/

O cérebro da transformação, com notebooks que executam cada etapa do pipeline, desde a ingestão até a camada gold.



Estrutura do Projeto: Scripts e Automação

01

01_ingestion_landing

Importa os arquivos CSV.

02

02_bronze_layer

Organiza dados brutos.

03

03_silver_layer

Limpa, corrige e relaciona dados.

04

04_prepare_to_gold

Prepara para análises.

05

05_gold_layer

Gera dados finais analíticos.

A pasta **jobs/** com pipeline_job.yaml indica a ordem de execução para automatizar tudo.

O que é o Databricks?

O Databricks é uma plataforma unificada de engenharia e análise de dados, escalável, rápida e ideal para construir e gerenciar pipelines de dados complexos.

- Permite rodar códigos em Python, SQL e Spark no mesmo ambiente.
- Seu Unity Catalog organiza e gerencia dados de forma centralizada, garantindo governança e controle.

Por que usamos neste projeto?

- Para executar o pipeline ETL completo (ingestão → bronze → silver → gold).
- Para transformar grandes volumes de dados com alta eficiência.
- Para armazenar tabelas processadas no Unity Catalog, eliminando a necessidade de exportação.
- Para garantir organização, controle de acesso e governança dos dados do mercado de trabalho.

Onde ele entra na arquitetura?

- É o núcleo central do processamento do pipeline de dados.
- Recebe os dados brutos e executa todas as camadas: Landing → Bronze → Silver → Gold.
- Armazena todas as tabelas analíticas finais no Unity Catalog.
- O Metabase (ferramenta de visualização) se conecta diretamente ao Databricks para acesso aos dados.

Metabase – Ferramenta de Visualização de Dados

O que é o Metabase?

- Uma ferramenta de **BI simples e intuitiva**
- Permite criar gráficos, tabelas e dashboards **sem necessidade de programação**
- Ótima para visualizar resultados de pipelines de dados

Por que usamos neste projeto?

- Para transformar os dados processados em **visualizações claras**
- Para analisar e contar a "história" dos dados
- Para facilitar a apresentação e interpretação dos resultados

Onde ele entra na arquitetura?

- **Depois** que os dados passam pelo ETL no Databricks
- Se conecta **diretamente** ao Databricks (Unity Catalog)
- Gera gráficos e dashboards a partir dessa conexão, sendo a **camada final de visualização**

Estrutura do Projeto: Documentação e Ambiente

docs/

Contém o arquivo architecture.png, que mostra um desenho da arquitetura do projeto e o caminho que os dados percorrem.

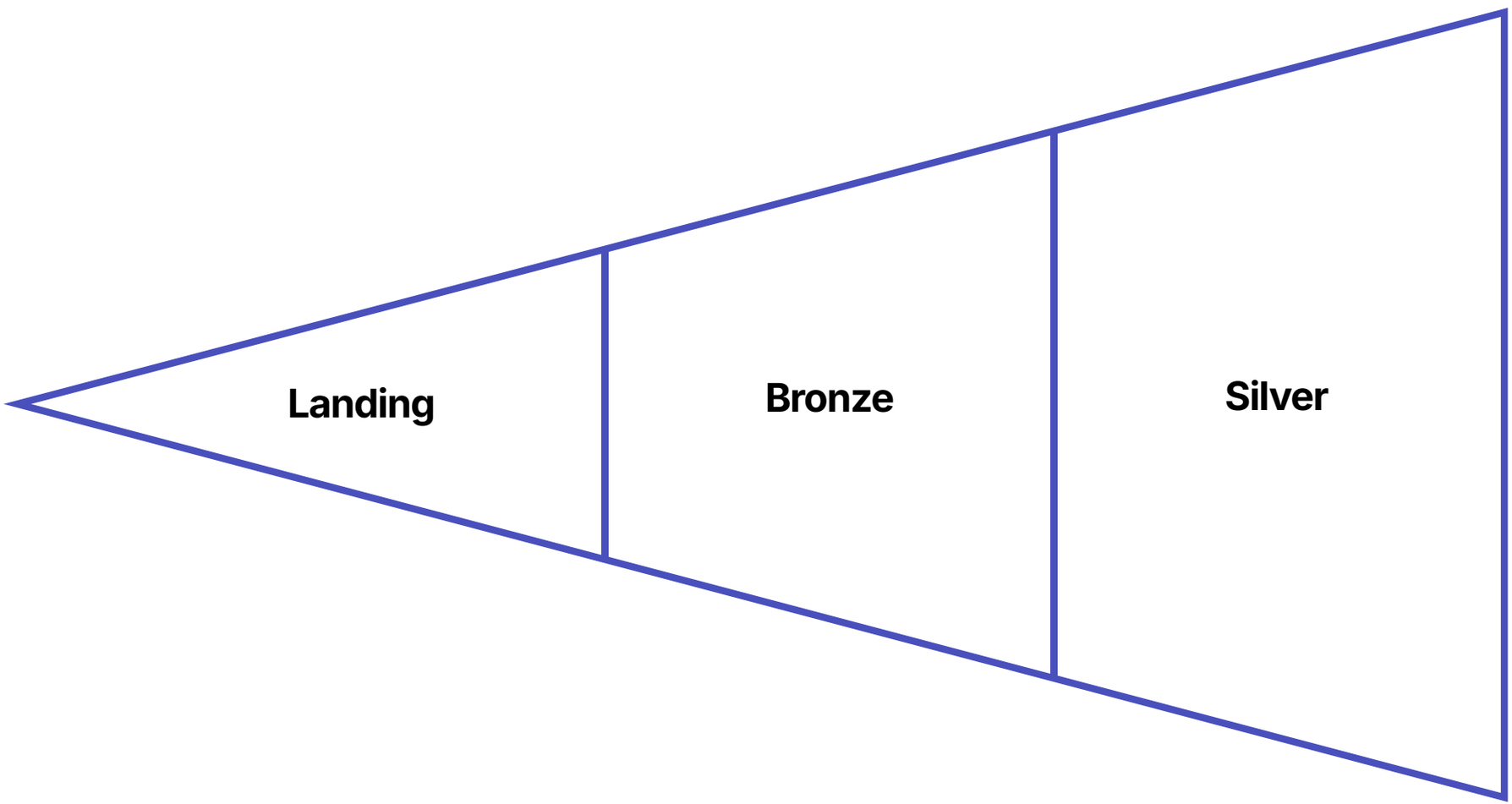
Arquivos de Ambiente

docker-compose.yml: Garante que o ambiente seja igual em qualquer computador.

pyproject.toml: Define as dependências do projeto (bibliotecas).

Arquitetura em Camadas

Os dados passam por etapas progressivas de refinamento.



Cada camada adiciona valor, transformando dados brutos em insights acionáveis.

Camadas de Transformação

Landing

Dados entram como vêm, sem alterações.

Bronze

Dados organizados e padronizados.

Silver

Dados limpos e conectados (ex: vagas + empresas).

Gold

Tabelas finais para análises (salários, skills, etc.).

O que o Projeto Entrega?

Dados limpos, integrados e organizados para dashboards, relatórios e análises.

Dados Unificados

Vagas e empresas.

Salários Confiáveis

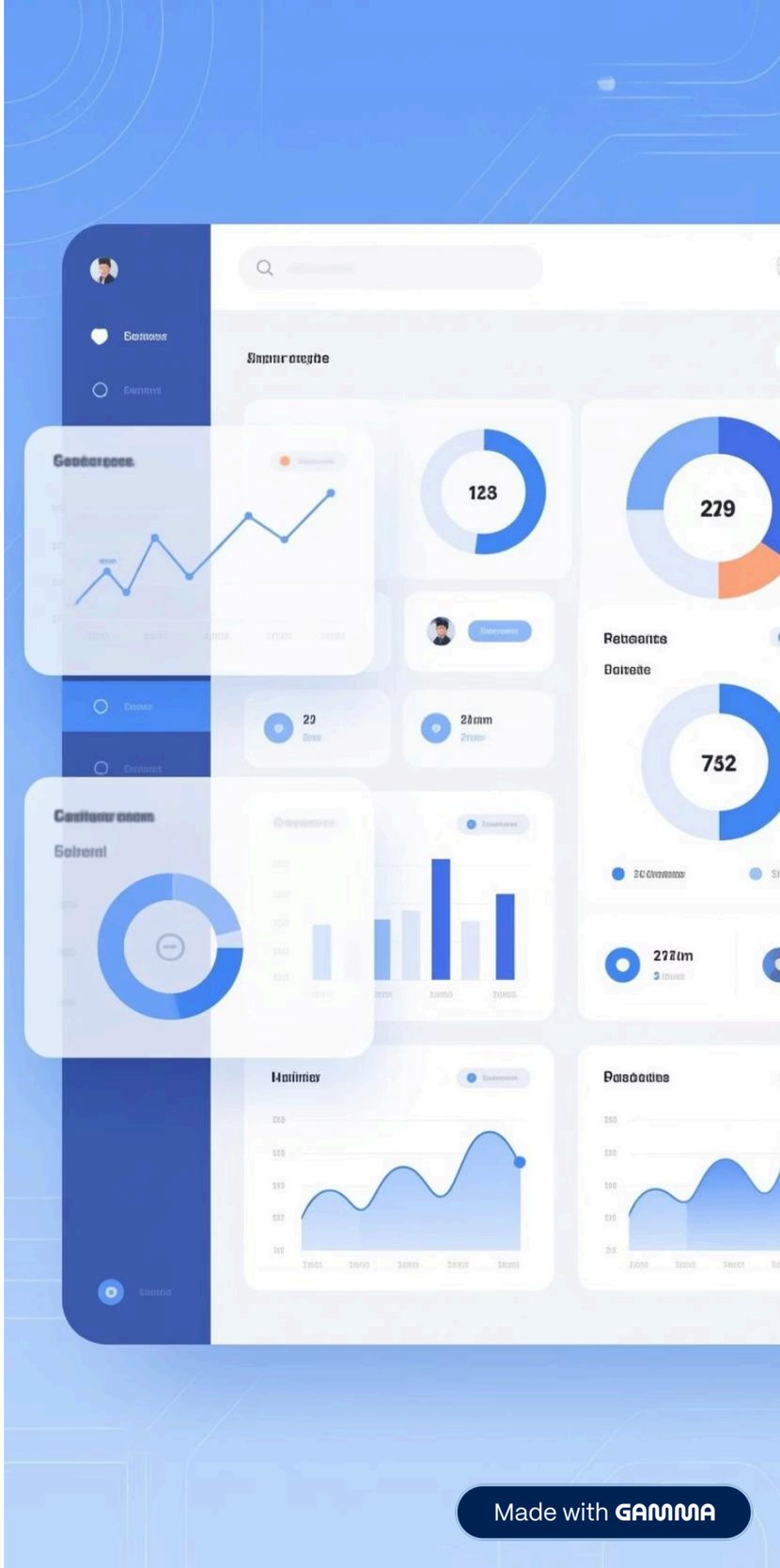
Informações precisas.

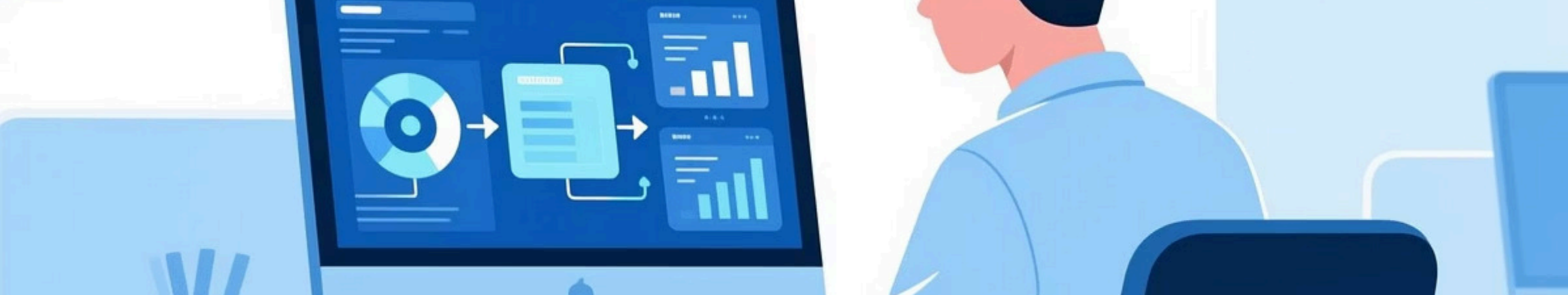
Skills Procuradas

Lista das mais relevantes.

Cruzamento

Empresas e vagas.





Por Que Isso é Útil?

Este projeto demonstra a construção de um pipeline ETL completo e profissional, capaz de transformar dados brutos em informação útil.

- Evita análises erradas devido a dados sujos.
- Automatiza a limpeza e integração dos dados.
- Facilita a criação de dashboards profissionais.
- Utilizado por engenheiros de dados no mundo real.

Acesse o Projeto no GitHub

Para uma visão detalhada do código-fonte, estrutura e documentação completa, explore o repositório do projeto no GitHub:

Ver o Repositório

Obrigado!

Esperamos que este projeto demonstre o poder da engenharia de dados para transformar informações brutas em inteligência de mercado acionável. Estamos à disposição para futuras colaborações e para explorar novos desafios no mundo dos dados.