

Documentação do Pipeline ETL - Teste Técnico Analista de Dados

Data: 22/04/2025

Documentação técnica do pipeline ETL desenvolvido para o teste de Analista de Dados.

Sumário

1. Introdução
2. Análise dos Dados
3. Transformações Realizadas
4. Estrutura do Banco de Dados
5. Implementação do Pipeline ETL
6. Resultados e Estatísticas
7. Instruções de Execução

1. Introdução

Este documento apresenta a documentação técnica do pipeline ETL (Extract, Transform, Load) desenvolvido como parte do teste técnico para a posição de Analista de Dados. O objetivo do projeto foi construir um pipeline ETL que lê dados de um arquivo Excel, realiza transformações necessárias e carrega corretamente essas informações em um banco de dados PostgreSQL. O pipeline foi implementado em Python, utilizando bibliotecas como pandas para manipulação de dados e psycopg2 para interação com o banco de dados PostgreSQL. O processo completo inclui extração de dados de múltiplas abas de uma planilha Excel, transformações para tratamento de valores nulos, formatação de documentos e padronização de campos, e finalmente a carga dos dados em tabelas relacionais no PostgreSQL.

2. Análise dos Dados

O arquivo Excel fornecido contém duas abas: 'clientes' e 'vendas'. A análise inicial dos dados revelou a estrutura e características de cada conjunto de dados.

2.1. Dados de Clientes

A aba 'clientes' contém 30 registros com informações como id_cliente, nome, email, documento, tipo_pessoa e tipo_contato. Durante a análise, identificamos que alguns campos apresentavam valores nulos, especialmente tipo_pessoa (60% dos registros) e tipo_contato (36.67% dos registros).

Amostra dos dados de clientes:

id_cliente	nome	email	documento	tipo_pessoa	tipo_contato
1	Bárbara Pires	gabriel34@araujo.org	43856291709	FISICA	email
2	João Lucas Alves	ocaldeira@hotmail.com	86.452.917/0001-42	nan	nan
3	Sr. Henrique Mendes	thales85@uol.com.br	12376954000182	nan	nan
4	Sra. Lara Castro	lopesalicia@castro.br	52609431000153	JURIDICA	email
5	Natália Carvalho	alice40@yahoo.com.br	91423680000101	nan	celular

2.2. Dados de Vendas

A aba 'vendas' contém 100 registros com informações como id_venda, id_cliente, data_venda e valor. Todos os campos estavam preenchidos, sem valores nulos. O valor das vendas varia de R\$ 70.99 a R\$ 999.25, com média de R\$ 516.45.

Amostra dos dados de vendas:

id_venda	id_cliente	data_venda	valor	mes	ano
1000	12	2024-12-02 00:00:00	112.65	12	2024
1001	16	2024-10-31 00:00:00	73.79	10	2024
1002	13	2024-11-16 00:00:00	567.57	11	2024
1003	22	2025-01-09 00:00:00	179.68	1	2025
1004	29	2024-12-04 00:00:00	304.88	12	2024

3. Transformações Realizadas

Com base na análise dos dados, foram identificadas e implementadas diversas transformações para melhorar a qualidade e consistência dos dados antes de carregá-los no banco de dados PostgreSQL.

3.1. Transformações nos Dados de Clientes

As seguintes transformações foram aplicadas aos dados de clientes:

- Tratamento de valores nulos em tipo_pessoa: preenchidos com 'FISICA' como valor padrão
- Padronização de tipo_pessoa: convertido para maiúsculas (FISICA, JURIDICA)
- Tratamento de valores nulos em tipo_contato: preenchidos com 'email' como valor padrão
- Padronização de tipo_contato: convertido para minúsculas (email, celular)
- Formatação de documentos: CPF formatado como 000.000.000-00 e CNPJ como 00.000.000/0000-00
- Adição de campo data_cadastro: preenchido com a data atual

3.2. Transformações nos Dados de Vendas

As seguintes transformações foram aplicadas aos dados de vendas:

- Arredondamento de valores para 2 casas decimais
- Extração de mês e ano da data de venda para facilitar análises temporais
- Adição de campo status_venda: todas as vendas foram marcadas como 'CONCLUÍDA'

4. Estrutura do Banco de Dados

O banco de dados PostgreSQL foi estruturado com duas tabelas principais: clientes e vendas. A estrutura foi projetada para manter a integridade referencial e facilitar consultas analíticas.

4.1. Tabela de Clientes

A tabela 'clientes' foi criada com a seguinte estrutura:

```
CREATE TABLE clientes ( id_cliente INTEGER PRIMARY KEY, nome VARCHAR(100) NOT NULL,
email VARCHAR(100) NOT NULL, documento VARCHAR(20), documento_formatado VARCHAR(20),
tipo_pessoa VARCHAR(10) NOT NULL, tipo_contato VARCHAR(10) NOT NULL, data_cadastro
DATE NOT NULL );
```

4.2. Tabela de Vendas

A tabela 'vendas' foi criada com a seguinte estrutura:

```
CREATE TABLE vendas ( id_venda INTEGER PRIMARY KEY, id_cliente INTEGER NOT NULL,
data_venda DATE NOT NULL, valor NUMERIC(10, 2) NOT NULL, mes_venda INTEGER NOT NULL,
ano_venda INTEGER NOT NULL, status_venda VARCHAR(20) NOT NULL, FOREIGN KEY
(id_cliente) REFERENCES clientes (id_cliente) );
```

5. Implementação do Pipeline ETL

O pipeline ETL foi implementado em Python, utilizando diversas bibliotecas para manipulação de dados e interação com o banco de dados. O script principal (etl.py) contém todas as funções necessárias para extrair, transformar e carregar os dados.

5.1. Processo de Extração (Extract)

A extração dos dados é realizada a partir do arquivo Excel, utilizando a biblioteca pandas. O script lê as abas 'clientes' e 'vendas' e carrega os dados em DataFrames para processamento.

5.2. Processo de Transformação (Transform)

A transformação dos dados inclui tratamento de valores nulos, padronização de campos, formatação de documentos e criação de campos derivados. Todas as transformações são realizadas utilizando funcionalidades do pandas.

5.3. Processo de Carga (Load)

A carga dos dados é realizada utilizando a biblioteca psycopg2 para conexão com o PostgreSQL. O script cria as tabelas no banco de dados (se não existirem) e insere os dados transformados. O processo de carga utiliza a cláusula ON CONFLICT para evitar duplicações em caso de reexecução do script.

6. Resultados e Estatísticas

Após a execução do pipeline ETL, os dados foram carregados com sucesso no banco de dados PostgreSQL. A seguir, apresentamos algumas estatísticas e insights obtidos a partir dos dados processados.

6.1. Estatísticas de Clientes

- Total de clientes: 30
- Clientes pessoa física: 8 (26.7%)
- Clientes pessoa jurídica: 4 (13.3%)
- Preferência de contato por email: 7 (23.3%)
- Preferência de contato por celular: 12 (40.0%)

6.2. Estatísticas de Vendas

- Total de vendas: 100
- Valor total das vendas: R\$ 51644.91
- Valor médio por venda: R\$ 516.45
- Valor mínimo de venda: R\$ 70.99
- Valor máximo de venda: R\$ 999.25

7. Instruções de Execução

Para executar o pipeline ETL, siga as instruções abaixo:

7.1. Requisitos

- Python 3.6+
- PostgreSQL 12+
- Bibliotecas Python: pandas, openpyxl, psycopg2, numpy

7.2. Execução do Script

Para executar o pipeline ETL, utilize o seguinte comando:

```
python etl.py
```