

Universidade Federal de Pernambuco

Centro de Informática - CIn

Registro das Infrações de Trânsito



Relatório do Projeto de Integração

Banco de dados - 2025.1

Adriana Theil Melcop Castro - atmcc
Eduarda Vitória Albuquerque Sales - evas
Gustavo Felipe Alves da Silva - gfas2
Júlia Zovka de Souza - jzs
Lucas Guimarães Fernandes - lgf
Marcela Pereira Raposo - mpr

1. Descrição da Base de Dados Unificada:

A base de dados unificada consolidou informações sobre registros de infrações de trânsito ocorridas na cidade de Recife, nos anos de **2023, 2024 e 2025**, oriundas de diferentes arquivos CSV, disponíveis no site da prefeitura do Recife. Após o processo de integração, os dados foram organizados em uma única tabela no PostgreSQL hospedado no Supabase.

Bases antes das transformações

1	<code>datainfracao</code>	<code>1064812 non-null object</code>
2	<code>horainfracao</code>	<code>1064812 non-null object</code>
3	<code>dataimplantacao</code>	<code>1064812 non-null object</code>
4	<code>agenteequipamento</code>	<code>1002028 non-null object</code>
5	<code>infracao</code>	<code>1064812 non-null int64</code>
6	<code>descricaoinfracao</code>	<code>1064812 non-null object</code>
7	<code>amparolegal</code>	<code>1064812 non-null object *</code>
8	<code>localcometimento</code>	<code>1064812 non-null object *</code>

A base de dados antes de ser tranformada era composta por essas colunas, a maioria tratada como objeto pelo pandas.

- Nas colunas `datainfracao` e `dataimplantacao` tinha uma inconsistência relacionada ao ano de 2024.Enquanto 2023 e 2025 demonstravam esse dado como `date`, 2024 era como `datetime`.
- A coluna `agenteequipamento` tinha muitas linhas sem valores preenchidos
- A coluna `amparolegal` tinha muitas linhas com informações sem sentido “SENTIDO OLINDA”

Atributos principais pós-modificação:

- `data_infracao`
- `hora_infracao`

- data_implementacao
- agente Equipamento
- cod_infracao
- descricao_infracao
- local_cometimento
- artigo
- subdivisao_artigo

Após o tratamento dos dados definimos que as colunas relacionadas a data ficariam em DATE, a de hora no formato TIME, cod_infracao em int e o restante em string. Fizemos também a padronização dos dados esperados em cada coluna. As linhas que tinham valores nulos ou que não faziam sentido para o dado armazenado na coluna foram trocadas por “Não informado”.

Metadados:

Metadados			
Coluna	Tipo	Restrições	Descrição
data_infracao	DATE	NOT NULL	Data que a infração foi cometida
hora_infracao	TIME	NOT NULL	Hora que a infração foi cometida
data_implantacao	DATE	NOT NULL	Data que a infração foi registrada no sistema
agente_equipamento	TEXT	NOT NULL	O meio ou equipamento utilizado pelo agente de trânsito para registrar a infração
cod_infracao	BIGINT	NOT NULL	Código que indentifica qual infração foi cometida
descricao_infracao	TEXT	NOT NULL	Descrição sobre a infração
local_cometimento	TEXT	NOT NULL	Local onde ocorreu a infração
artigo	TEXT	NOT NULL	O número do artigo do CTB que foi infringido
subdivisao_artigo	TEXT	NOT NULL	Inciso ou parágrafo dentro de um artigo do Código de Trânsito Brasileiro

2. Explicação Detalhada do Processo de Integração:

Passo 1: Coleta de dados

- Os dados foram obtidos a partir de três arquivos CSV, cada um correspondente a um ano (2023, 2024 e 2025), sobre registros de infrações de trânsito. Baixamos os arquivos e alocamos eles na pasta 'datasets' do repositório.

Os passos abaixo mudam de ordem de acordo com ETL e ELT:

Passo 2: Transformação

- As colunas foram tratadas: foram padronizados nomes de colunas, adicionamos o campo subdivisao_artigo, tratamos de valores que não faziam sentido e valores nulos e formatamos os tipos de dados das colunas.

Passo 3: Criação da tabela unificada

- Depois dessas transformações temos uma base mais uniformizada e amigável a consultas, tanto na parte de ETL e ELT.

3. Justificativa da Escolha da Base de Dados:

A escolha pela base de infrações de trânsito se deu por diversos motivos:

- Disponibilidade pública e acessível de dados abertos
- Relevância social e urbana: permite identificar comportamentos de risco e padrões de infração
- Permite análises práticas com impacto real, como bairros com mais infrações ou horários críticos.

Somado a esses motivos, diante de uma análise rápida, a base já parecia ter um certo grau de uniformidade, com número de colunas e tipos de dados iguais nos anos escolhidos para análise.

Além disso, o uso do **Supabase com PostgreSQL** foi motivado pela facilidade de

integração, visualização rápida e acesso remoto ao banco.

4. Descrição dos Processos de Transformação Aplicados:

Como as colunas tinham o mesmo número de colunas, e aparentemente o mesmo tipo de dado, decidimos concatenar as três em uma única tabela no formato csv. Ao analisar de fato a estrutura do arquivo percebemos que existiam umas inconsistências apesar de as informações com `df.info()` estarem iguais. Nesse sentido, tinham valores nulos em uma das colunas de 2025 e a coluna "datainfracao" estava diferente nos arquivos. Diante disso realizamos as seguinte transformações antes de poder concatenar os arquivos.

- **Padronização das datas:** em uma inspeção mais detalhada do csv final percebemos que o formato está diferente nos dados referentes ao ano de 2024. 2023 = "2023-10-26;", 2024 = "06/05/2024 00:00", 2025 = "2025-04-22". Por isso, transformamos todas em DATE, tanto no ETL, quanto no ELT.

Diante do arquivo concatenado aplicamos o restante das transformações

- **Renomeação das colunas:** 'datainfracao' para 'data_infracao', todas as colunas eram nomes concatenados, fizemos essa renomeação em todas para facilitar a leitura e entendimento.
 - datainfracao -> data_infracao
 - horainfracao -> hora_infracao
 - dataimplementacao -> data_implementacao
 - agenteequipamento -> agente_equipamento
 - infracao -> cod_infracao
 - descricaoinfracao -> descricao_infracao
 - localcometimento -> local_cometimento
- **Divisão da coluna amparo legal:** Decidimos dividir a coluna em artigo e subdivisao_artigo para facilitar uma consulta sobre os incisos violados.
- **Corrigindo valores inválidos na coluna artigo:** existiam valores errados para o artigo. Valores como: 'SENTIDO OLINDA', 'ILHA DO LEITE', 'SENTIDO CIDADE/SUBURBIO.'e 'Art. 244 inciso VIII' - esse último pelo fato de não ter o separador que usamos na criação da coluna amparo legal. Então corrigimos as linhas com outros problemas substituindo por 'Não Informado' e separamos 'Art. 244 inciso VIII' e outras ocorrências desse tipo nas colunas correspondentes, artigo e subdivisão artigo.
- **Padronização dos valores em subdivisao_artigo:** inciso -> Inc. inc. -> Inc. parágrafo único -> § único valores estranhos ou nulos -> Não Informado
- **Tratamento de valores nulos ou inconsistentes:** a coluna agente_equipamento

do ano de 2025 tinha 62784 valores nulos, como era um valor pequeno em comparação com a soma de linhas dos três anos decidimos manter essa coluna e substituir por “não informado”, como fizemos em outras 5 colunas: agente Equipamento, descricao_infracao, local_cometimento, artigo e subdivisao_artigo

- **Conversão de tipos:** Duas colunas foram convertidas para o tipo DATE e uma para o tipo TIME, visando facilitar futuras consultas. A coluna "artigo", embora contenha números, foi mantida como STRING. Dessa forma, quando um campo estiver vazio, ele exibirá "Não informado" em vez de "0", mantendo a consistência com as demais colunas e evitando confusão com um possível Artigo 0.
 - dft['data_infracao'] -> DATE
 - dft['hora_infracao'] -> TIME
 - dft['data_implantacao'] -> DATE
 - dft['agente_equipamento'] -> STRING
 - dft['descricao_infracao'] -> STRING
 - dft['local_cometimento'] -> STRING
 - dft['artigo'] -> STRING
 - dft['subdivisao_artigo'] -> STRING

5. Comparativo entre ETL e ELT:

ETL (Extract, Transform, Load): Transformações feitas antes de carregar no banco

Vantagens:

- Útil quando o ambiente de destino tem capacidade limitada
- Mais controle durante o pré-processamento

Desvantagens:

- A análise de dados só fica disponível após o processo de transformação, o que demora

ELT (Extract, Load, Transform): Transformações feitas após carregar tabela no banco de dados.

Vantagens:

- Os dados são carregados "crus" no banco e transformados ali mesmo
- Tira proveito da performance do banco (Supabase, no nosso caso)

- Facilita reprocessamentos e pipelines dinâmicos

Desvantagens:

- Só é viável em plataformas de dados que possuem um bom poder de processamento para executar transformações em larga escala de forma eficiente
- Custoso: paga pelo armazenamento usado

No nosso projeto:

- Utilizamos **ETL para as transformações iniciais** (tratamento, limpeza)
- E também **ELT para análises mais pesadas via SQL no Supabase**

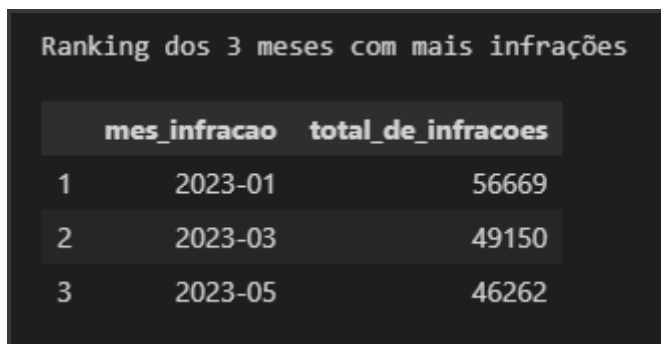
→ Isso nos permitiu **testar os dois modelos**, entendendo as vantagens de cada um no contexto real de um pipeline híbrido.

6. Apresentação de Três Análises e Insights:

Análise 1: Mês com maior número de infrações

- Insight:

ELT:



Ranking dos 3 meses com mais infrações

	mes_infracao	total_de_infracoes
1	2023-01	56669
2	2023-03	49150
3	2023-05	46262

Análise 2: Horários com mais registros

- Insight:

ELT:

Ranking dos 3 horários com mais registros

	hora_do_dia	total_de_infracoes
1	15.0	112099
2	9.0	104528
3	10.0	101831

Análise 3: Tipo de infração mais comum

- Insight:

Ranking dos 3 tipos de infração mais comuns:

	descricao_infracao	total_de_ocorrencias
1	Transitar em velocidade superior à máxima perm...	306426
2	Estacionar o veículo em desacordo com as condi...	143327
3	Transitar na faixa ou via exclusiva regulam. ...	131447

7. Reflexão sobre o Aprendizado

Este projeto proporcionou um aprendizado significativo sobre integração de dados na prática. Alguns pontos marcantes:

- **Desafio:** Configuração inicial do ambiente no Supabase e lidar com inconsistências entre os arquivos dos diferentes anos. Entendimento de como usar pandas para manipular os datasets.
- **Superação:** Compreensão do fluxo de um pipeline de dados completo e uso real de ferramentas como o dbt para organizar a camada transformacional
- **Lição:** A importância de pensar na **qualidade dos dados desde o início** e a vantagem de adotar boas práticas de organização, versionamento e padronização

Além disso, foi um exercício de colaboração e aplicação prática de teorias aprendidas na disciplina, reforçando o valor de projetos que simulam o mercado real.