

Case de QA e Análise de Dados – GA4 & BigQuery Simulation (MySQL Local)

Este projeto simula um cenário real de qualidade de dados e análise de conversão por canal utilizando dados exportados do Google Analytics 4 (GA4) para BigQuery, mas recriados em MySQL local para evitar custos de cloud.

O case foi estruturado para demonstrar habilidades de QA de dados, ETL, validação e análise de impacto da qualidade.

1 Estrutura do Projeto

CASE/

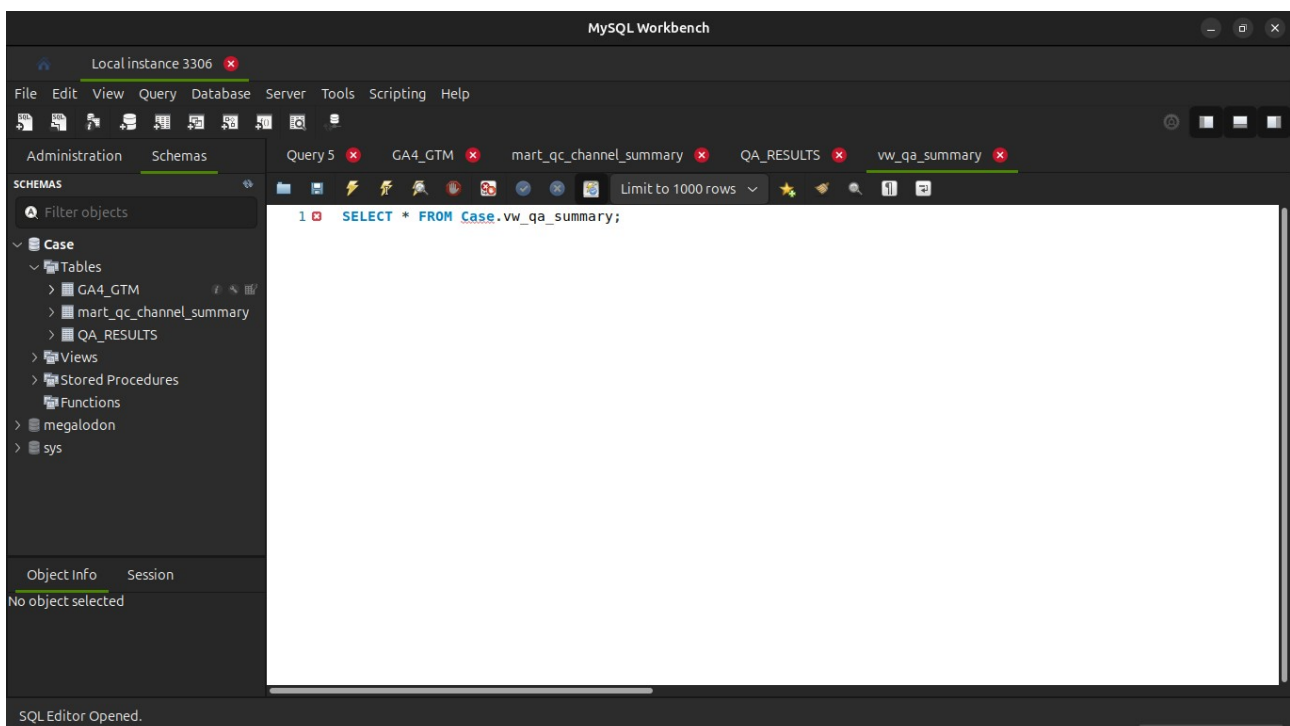
- CSV/ # Exportações CSV para uso no Power BI
- SQL/ # Scripts SQL organizados por etapa
- Power_BI/ # Relatórios e arquivos de conexão
- venv/ # Ambiente virtual Python
- .env # Variáveis de ambiente (credenciais MySQL)
- populate_table.py # Script para popular dados simulados
- README.md # Documentação do projeto

2 Etapas do Desenvolvimento

1. Criação do Schema e Tabela de Eventos

Arquivo: Cria_Schema_e_tabela_GA4_GTM.sql

- Cria o schema Case e a tabela GA4_GTM que receberá os eventos GA4 simulados.
- Estrutura compatível com colunas comuns no GA4 Export.



2. População de Dados com Simulação de Problemas de Qualidade

Arquivo: populate_table.py

- Script Python que insere 300 registros na tabela, sendo 50 com problemas de qualidade (nulos, duplicatas, datas futuras, valores inválidos etc.).

- Utiliza variáveis de ambiente no .env para conexão MySQL.
- Garante aleatoriedade controlada para reproduzibilidade.

The screenshot shows an IDE with a Python script named `populate_table.py` and its execution output in the terminal.

Script Content:

```

181 def main():
182     charset="utf8mb4", autocommit=False
183 )
184 try:
185     with conn.cursor() as cur:
186         # (Opcional) limpar antes:
187         cur.execute(f"TRUNCATE TABLE `{MYSQL_DB}`.`{MYSQL_TABLE}`;")
188         cur.executemany(INSERT_SQL, rows)
189     conn.commit()
190     print(f"✓ Inseridos {len(rows)} registros em `{MYSQL_DB}`.`{MYSQL_TABLE}`")
191     print(f"⚠ Desse, {PROBLEMAS_QUALIDADE} possuem problemas de qualidade (nulos, duplicatas, futu
192 except Exception as e:
193     conn.rollback()
194     print("✗ Erro ao inserir registros:", e)
195     raise
196 finally:

```

Terminal Output:

```

shurillo@shurillo-IdeaPad-3-15IML05:~/Case$ source venv/bin/activate
(venv) shurillo@shurillo-IdeaPad-3-15IML05:~/Case$ /home/shurillo/Case/venv/bin/python /home/shurillo/Case/populat
e table.py
✓ Inseridos 300 registros em `Case`.`GA4 GTM`
⚠ Desse, 50 possuem problemas de qualidade (nulos, duplicatas, futuros etc.).
(venv) shurillo@shurillo-IdeaPad-3-15IML05:~/Case$

```

The screenshot shows MySQL Workbench with a query executed on the `GA4_GTM` table. The query is `SELECT * FROM Case.GA4_GTM;`. The result shows 12 rows of data.

#	id	event_id	event_ts	event_date	user_pseudo_id	session_id	event_name	source_medium	page_location
1	1	e4729480b9e14271	2025-08-09 16:49:06.823990	2025-08-09	user_4fmyzy	2169	purchase	email / mkt	https://site.test/pdt
2	2	c33b4daba0d145cf	2025-08-07 09:18:06.823990	2025-08-07	user_32oc6u	6155	add_to_cart	email / mkt	http://example.com
3	3	c6054fb722ab49d3	2025-08-07 11:42:06.823990	2025-08-07	user_0t0pvn	3266	add_to_cart	email / mkt	http://site.test/hom
4	4	ff854baf6e80401e	2025-08-14 10:02:06.825123	2025-08-14	user_t84azy	5371	random_event	unknown / ???	:/no-scheme
5	5	88148bd9c10746ac	2025-08-10 04:37:06.823990	2025-08-10	user_5jsq65	5889	add_to_cart	direct / (none)	https://site.test/plp
6	6	04c7717ae44e4682	2025-08-08 20:44:06.823990	2025-08-08	user_a753lc	2290	page_view	referral / partner	http://example.com
7	7	73ee05add5064fad	2025-08-07 00:34:06.823990	2025-08-07	user_j5ph0	4295	add_to_cart	direct / (none)	http://site.test/cont
8	8	fbe7ad13539e4ff4	2025-08-06 18:33:06.823990	2025-08-06	user_icatva	1964	purchase	referral / partner	https://example.co
9	9	851a5ccff84b4360	2025-08-06 20:05:06.823990	2025-08-06	user_05uiro	2545	purchase	facebook / cpc	http://example.com
10	10	a601c7c1086845dc	2025-08-10 05:27:06.823990	2025-08-10	user_m5igqp	5563	page_view	google / cpc	https://shop.local/h
11	11	dcbe922711ee4aa3	2025-08-06 14:46:06.823990	2025-08-06	user_74frho	1960	page_view	referral / partner	http://example.com
12	12	2545e2af5ab0546dc	2025-08-10 01:08:06.823990	2025-08-10	user_80u7fk	1958	purchase	email / mkt	https://site.test/abc

Problemas simulados:

- source_medium nulo ou inválido
- event_id duplicado
- Datas futuras em event_ts
- event_name fora do domínio permitido
- page_location malformatado
- Compras (purchase) sem transaction_id ou valores inválidos
- items negativos ou irreais
- session_id nulo

3. Validação de Qualidade dos Dados

Arquivo: Cria_Procedure_de_validacao_dedados.sql

- Procedure sp_run_quality_checks que:
- Executa regras de validação nos campos-chave.
- Registra resultados na tabela QA_RESULTS.
- View vw_qa_summary para resumo da qualidade dos dados.

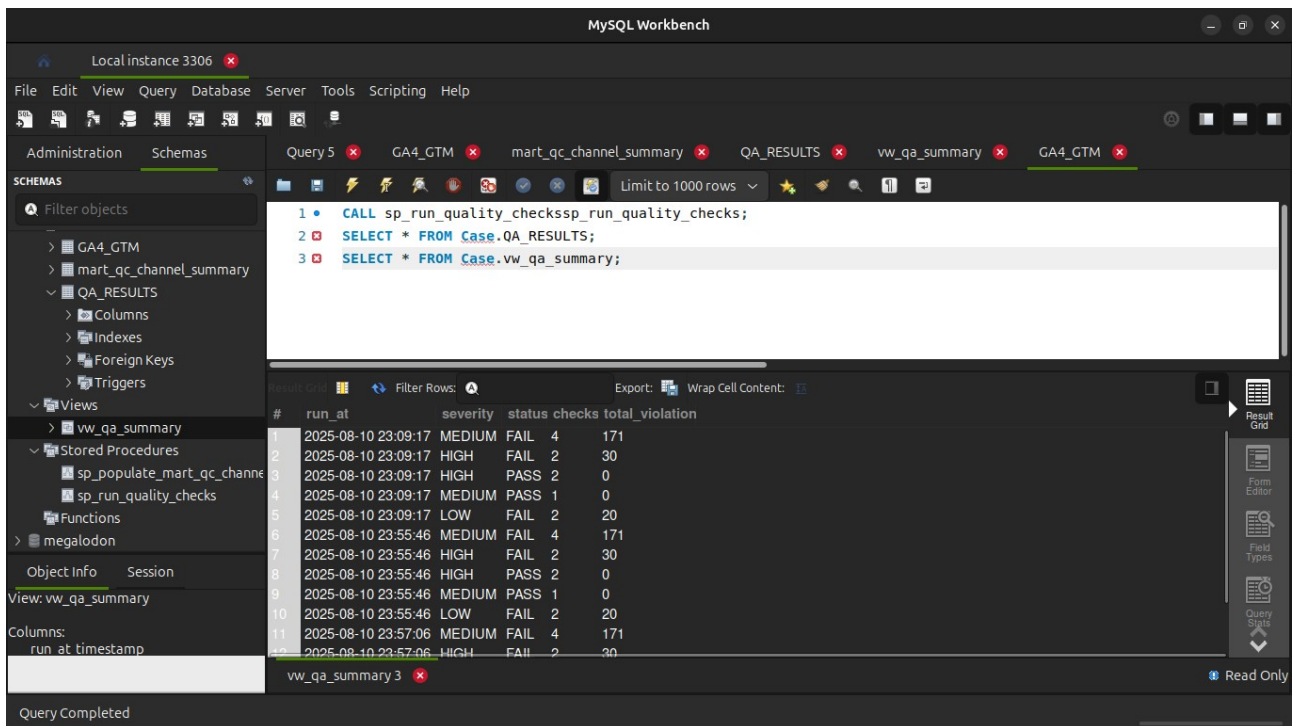
The screenshot shows the MySQL Workbench interface. The left sidebar displays the 'SCHEMAS' tree with the 'Case' database selected. The 'QA_RESULTS' table is highlighted under the 'Tables' section. The main query editor shows the following SQL code:

```
1 CALL sp_run_quality_checks;
2 SELECT * FROM Case.QA_RESULTS;
```

The 'Results' tab at the bottom displays the output of the query, showing 12 rows of data. The columns are: #, id, run_at, check_name, description, severity, threshold_violation, violation, status, and sample. The data indicates various quality issues, such as null values, future dates, and invalid event names.

#	id	run_at	check_name	description	severity	threshold_violation	violation	status	sample
1	1	2025-08-10 23:09:17	NULL_source_medium	Eventos com source_medium nulo	MEDIUM	0	31	FAIL	[[{"id": 1, "source_medium": null, "event_ts": "2025-08-10 23:09:17", "event_name": "FUTURE_events", "page_location": "http://www.example.com", "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]
2	2	2025-08-10 23:09:17	FUTURE_events	event_ts > NOW() + 24h	HIGH	0	20	FAIL	[[{"id": 2, "source_medium": "GA4_GTM", "event_ts": "2025-08-11 00:00:00", "event_name": "FUTURE_events", "page_location": "http://www.example.com", "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]
3	3	2025-08-10 23:09:17	INVALID_event_name	event_name fora do domínio permit...	MEDIUM	0	50	FAIL	[[{"id": 3, "source_medium": "GA4_GTM", "event_ts": "2025-08-10 23:09:17", "event_name": "INVALID_event_name", "page_location": "http://www.example.com", "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]
4	4	2025-08-10 23:09:17	DUP_event_id	Registros excedentes por event_id ...	HIGH	0	10	FAIL	[[{"id": 4, "source_medium": "GA4_GTM", "event_ts": "2025-08-10 23:09:17", "event_name": "FUTURE_events", "page_location": "http://www.example.com", "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]
5	5	2025-08-10 23:09:17	PURCHASE_missing_txid	Eventos purchase sem transaction_id	HIGH	0	0	PASS	[[{"id": 5, "source_medium": "GA4_GTM", "event_ts": "2025-08-10 23:09:17", "event_name": "FUTURE_events", "page_location": "http://www.example.com", "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]
6	6	2025-08-10 23:09:17	PURCHASE_bad_value	purchase com value NULL/<=0 ou ...	HIGH	0	0	PASS	[[{"id": 6, "source_medium": "GA4_GTM", "event_ts": "2025-08-10 23:09:17", "event_name": "FUTURE_events", "page_location": "http://www.example.com", "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]
7	7	2025-08-10 23:09:17	INVALID_currency	currency não está na whitelist	MEDIUM	0	50	FAIL	[[{"id": 7, "source_medium": "GA4_GTM", "event_ts": "2025-08-10 23:09:17", "event_name": "FUTURE_events", "page_location": "http://www.example.com", "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]
8	8	2025-08-10 23:09:17	PURCHASE_bad_items	purchase com items NULL/<=0/>1...	MEDIUM	0	0	PASS	[[{"id": 8, "source_medium": "GA4_GTM", "event_ts": "2025-08-10 23:09:17", "event_name": "FUTURE_events", "page_location": "http://www.example.com", "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]
9	9	2025-08-10 23:09:17	NULL_page_location	page_location nulo	LOW	0	10	FAIL	[[{"id": 9, "source_medium": "GA4_GTM", "event_ts": "2025-08-10 23:09:17", "event_name": "FUTURE_events", "page_location": null, "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]
10	10	2025-08-10 23:09:17	MALFORMED_page_location	page_location não inicia com http://...	MEDIUM	0	40	FAIL	[[{"id": 10, "source_medium": "GA4_GTM", "event_ts": "2025-08-10 23:09:17", "event_name": "FUTURE_events", "page_location": "http://www.example.com", "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]
11	11	2025-08-10 23:09:17	NULL_session_id	session_id nulo	LOW	0	10	FAIL	[[{"id": 11, "source_medium": "GA4_GTM", "event_ts": "2025-08-10 23:09:17", "event_name": "FUTURE_events", "page_location": "http://www.example.com", "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]
12	12	2025-08-10 23:55:46	NULL_source_medium	Eventos com source_medium nulo	MEDIUM	0	31	FAIL	[[{"id": 12, "source_medium": null, "event_ts": "2025-08-10 23:55:46", "event_name": "FUTURE_events", "page_location": "http://www.example.com", "purchase": "2025-08-10 23:09:17", "items": 10, "session_id": null}]]

The status bar at the bottom indicates 'Query Completed'.



Como executar:

```
CALL `Case`.`sp_run_quality_checks`();
```

-- Ver resultados detalhados

```
SELECT *
FROM `Case`.`QA_RESULTS`
WHERE run_at = (SELECT MAX(run_at) FROM `Case`.`QA_RESULTS`)
ORDER BY severity DESC, violations DESC;
```

-- Ver resumo

```
SELECT *
FROM `Case`.`vw_qa_summary`
WHERE run_at = (SELECT MAX(run_at) FROM `Case`.`QA_RESULTS`);
```

4. Cálculo de Conversão por Canal com e sem Qualidade

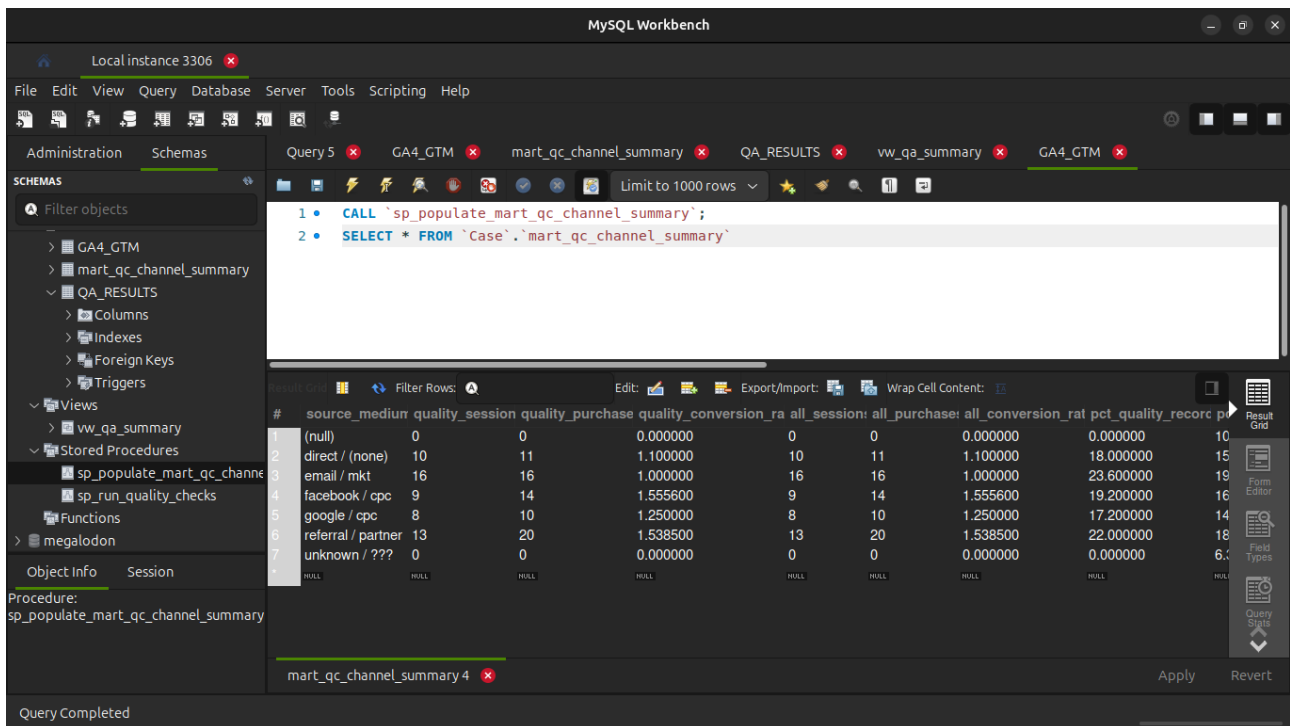
Arquivo: Cria_procedure_e_popula_calculos_conversao.sql

- Procedure sp_populate_mart_qc_channel_summary:
- Calcula taxa de conversão apenas com dados de qualidade assegurada.
- Calcula taxa de conversão com todos os dados (incluindo problemas).
- Mostra impacto (%) da falta de qualidade.
- Registra contagem de registros usados em cada cenário.

Como executar:

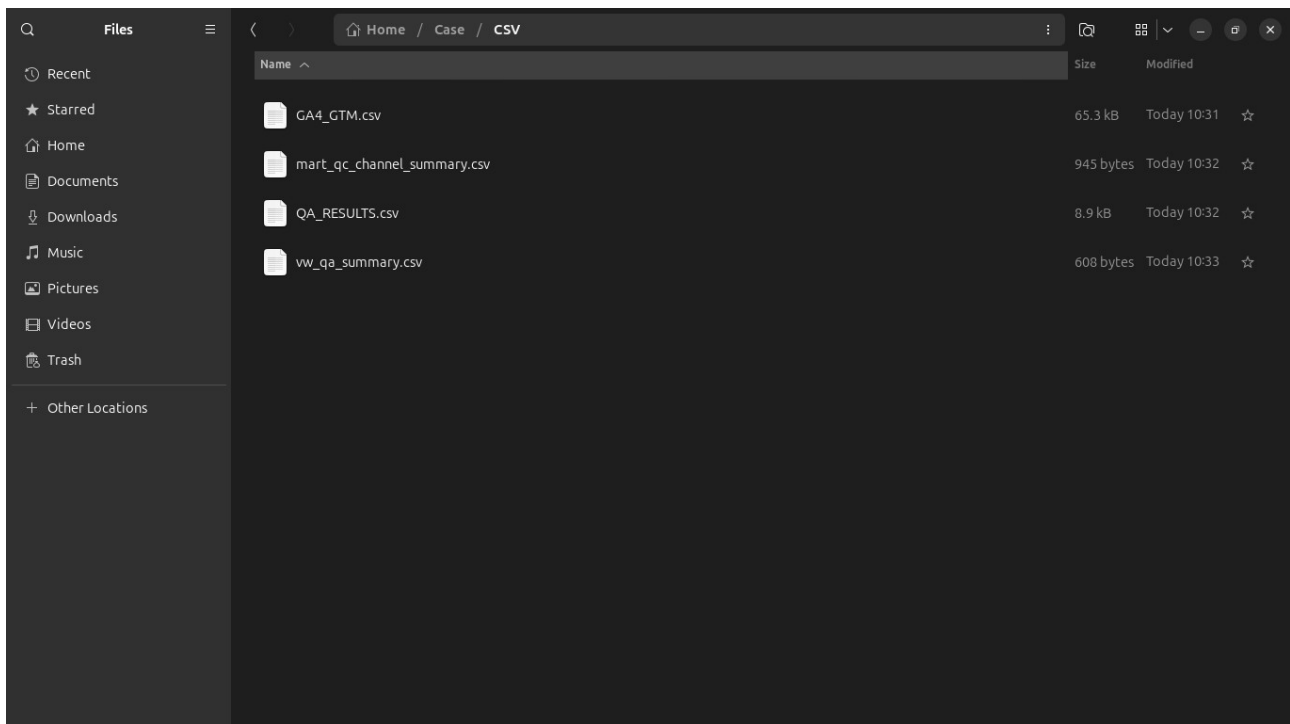
```
CALL `Case`.`sp_populate_mart_qc_channel_summary`();
```

```
SELECT *
FROM `Case`.`mart_qc_channel_summary`
ORDER BY all_sessions DESC, source_medium;
```



5. Exportação para Power BI (Passo extra)

- Exportação do MySQL para CSV (pasta /CSV).
- Desenvolvimento de dashboards no Power BI no Windows usando os arquivos exportados do Ubuntu.
- Recentemente migrei de sistema operacional e passei a utilizar Ubuntu, e, por não ter aprendido ainda como instalar o Power BI no Ubuntu, decidi extrair para CSV e desenvolver no ambiente windows o dashboard. Porém num cenário comum o BI seria conectado diretamente no BigQuery do GCP.



3 Ferramentas Utilizadas

- ChatGPT Plus – Apoio na concepção e automação
- MySQL Workbench 8.0 – Modelagem e execução SQL
- Python 3.13 – Bibliotecas PyMySQL e python-dotenv
- Ubuntu – Ambiente principal de desenvolvimento
- Windows 11 – Ambiente para criação de dashboards no Power BI



Desenvolvimento do BI no Power BI

Esta etapa teve como objetivo construir uma análise visual a partir dos dados simulados, permitindo identificar e mensurar o impacto de problemas de qualidade nos eventos registrados.

1 Conexão e Tratamento de Dados

- Conexão inicial realizada a partir de arquivos CSV, simulando uma conexão direta com o banco de dados.
- Conversão e tratamento dos tipos de dados de cada tabela para garantir consistência nas análises.
- Criação da tabela dCalendario, utilizada como dimensão de tempo para facilitar análises temporais.
- Criação de uma tabela chamada "Medidas" para organizar e centralizar todas as medidas DAX criadas.

2 Modelagem e Relacionamentos

- Estabelecimento dos relacionamentos entre as tabelas de fatos e a dCalendario, garantindo integridade na análise temporal.
- Relacionamento entre as tabelas de eventos (GA4_GTM) e a tabela de resultados de QA (QA_RESULTS) para cruzamento das informações de qualidade.

3 Medidas DAX Implementadas

- Quantidade de registros totais:
 $\text{Qtd. Registros} = \text{DISTINCTCOUNT}(\text{GA4_GTM}[\text{id}])$
- Quantidade de registros sem qualidade:
 $\text{Qtd. Registros bad} = \text{DISTINCTCOUNT}(\text{QA_RESULTS}[\text{id}])$
- Percentual de registros sem qualidade:
 $\% \text{ s/ Qualidade} = \text{DIVIDE}([\text{Qtd. Registros bad}], [\text{Qtd. Registros}])$

4 Visualizações Criadas

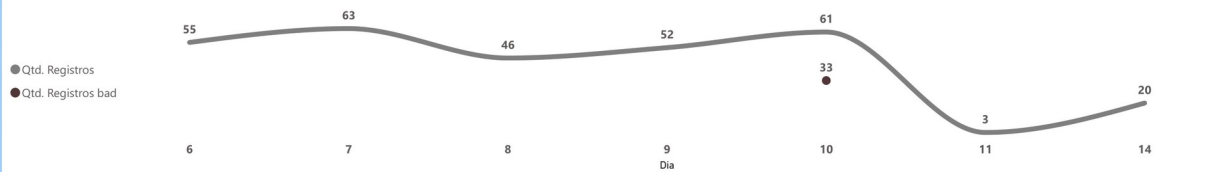
- Visão Geral: Cartões com KPIs de volume total de registros, volume de registros com problema e percentual de dados sem qualidade.
- Análise Temporal: Gráficos lineares e de colunas para acompanhar a evolução da qualidade dos dados ao longo do tempo.
- Impacto por Canal/Fonte: Visual comparando o volume e percentual de dados com problema por source_medium.

Qtd. Registros
300

Qtd. Registros bad
33

% s/ Qualidade
11,00%

Entrada de registros no banco (por dia)



Qtd. de Registros por evento

Evento	Qtd. Registros
purchase	71
add_to_cart	65
begin_checkout	58
page_view	56
random_event	11
view_page	11
Total	300

Detalhamento de problemas de qualidade

Descrição inconsistência	Nível inconsistência	Status	Qtd. Registros bad	% s/ Qualidade
currency não está na whitelist	MEDIUM	FAIL	3	1,00%
event_name fora do domínio permitido	MEDIUM	FAIL	3	1,00%
event_ts > NOW() + 24h	HIGH	FAIL	3	1,00%
Eventos com source_medium nulo	MEDIUM	FAIL	3	1,00%
Eventos purchase sem transaction_id	HIGH	PASS	3	1,00%
page_location não inicia com http:// ou https://	MEDIUM	FAIL	3	1,00%
page_location nulo	LOW	FAIL	3	1,00%
Total			33	11,00%