MVP – Engenharia de Dados

Sprint: Engenharia de Dados (40530010057_20230_01)

Rio de Janeiro, 01 de outubro de 2023

PUC-Rio – Pontifícia Universidade Católica do Rio de Janeiro
R. Marquês de São Vicente, 225 - Gávea, Rio de Janeiro – RJ

Curso: Pós-Graduação CIÊNCIA DE DADOS E ANALYTICS
Estudante: IVAN RIBEIRO TEIXEIRA
Matrícula: 4052023001146

Sprint: Engenharia de Dados (40530010057_20230_01)
MVP (minimum viable product)
Prof. Dr. Sérgio Lifschitz

1 –Introdução e Busca pelos Dados

Este trabalho reflete e consolida os ensinamentos e informações do Sprint: Engenharia de Dados (40530010057_20230_01) do Curso: Pós-Graduação CIÊNCIA DE DADOS E ANALYTICS, no formato de um MVP (minimum viable product), de modo a avaliar a capacidade de construir um pipeline de dados utilizando tecnologias na nuvem. O pipeline irá envolver a busca, coleta, modelagem, carga e análise dos dados.

2 - Objetivo

O dataset escolhido foi do seriado americano SuperLoja, ("SuperStore" lançado em 2015 e transmitida no Brasil por alguns canais fechados de streaming no Amazon Prime Vídeo e Netflix), disponível no portal público do Tableau (https://public.tableau.com/app/learn/sample-data).

A escolha de tal dataset deve-se: a) por se tratar de informações de vendas, área de negócio que já trabalho e domínio a mais de 20 anos e b) por ser uma série que assisti e gostei muito.

Com base no dataset "SuperLoja", o objetivo é responder algumas perguntas sobre as vendas de produtos e gerar alguns relatórios para tomada de decisão.

2.1 – Problema fictício a ser respondido

A Diretoria da SuperLoja, quer fazer alguns levantamentos para tomada de decisão de abertura de novas lojas em outras localidades. Para resolver este problema, foram feitas as seguintes perguntas para análise:

a) Análise 1: Qual é o Total de Vendas por Categoria de Material
b) Análise 2: Mostre o Lucro Total por Região
c) Análise 3: Qual a Média de Desconto por Segmento de Clientes
d) Análise 4: Faça uma Consulta que mostre as Vendas por Mês
e) Análise 5: Qual o Lucro por Material
f) Análise 6: Qual é a Quantidade de Pedidos por País
g) Análise 7: Faça uma consulta que mostre o Top 10 Clientes com Maior Valor de Vendas
h) Análise 8: Mostre o Total de Vendas por Ano e Segmento de Clientes

Diante exposto, foram elaboradas consultas SQL para responder as análises acima.

3 - Coleta dos dados

Como já informado, utilizei o dataset público do Seriado Televisio "Super loja", disponível no site do Tableau Público. Este dataset contém informações sobre produtos, vendas e lucros de uma empresa fictícia baseada em uma série de TV.

O dataset está disponível no seguinte link:

https://public.tableau.com/app/learn/sample-data

Dicionário de Dados

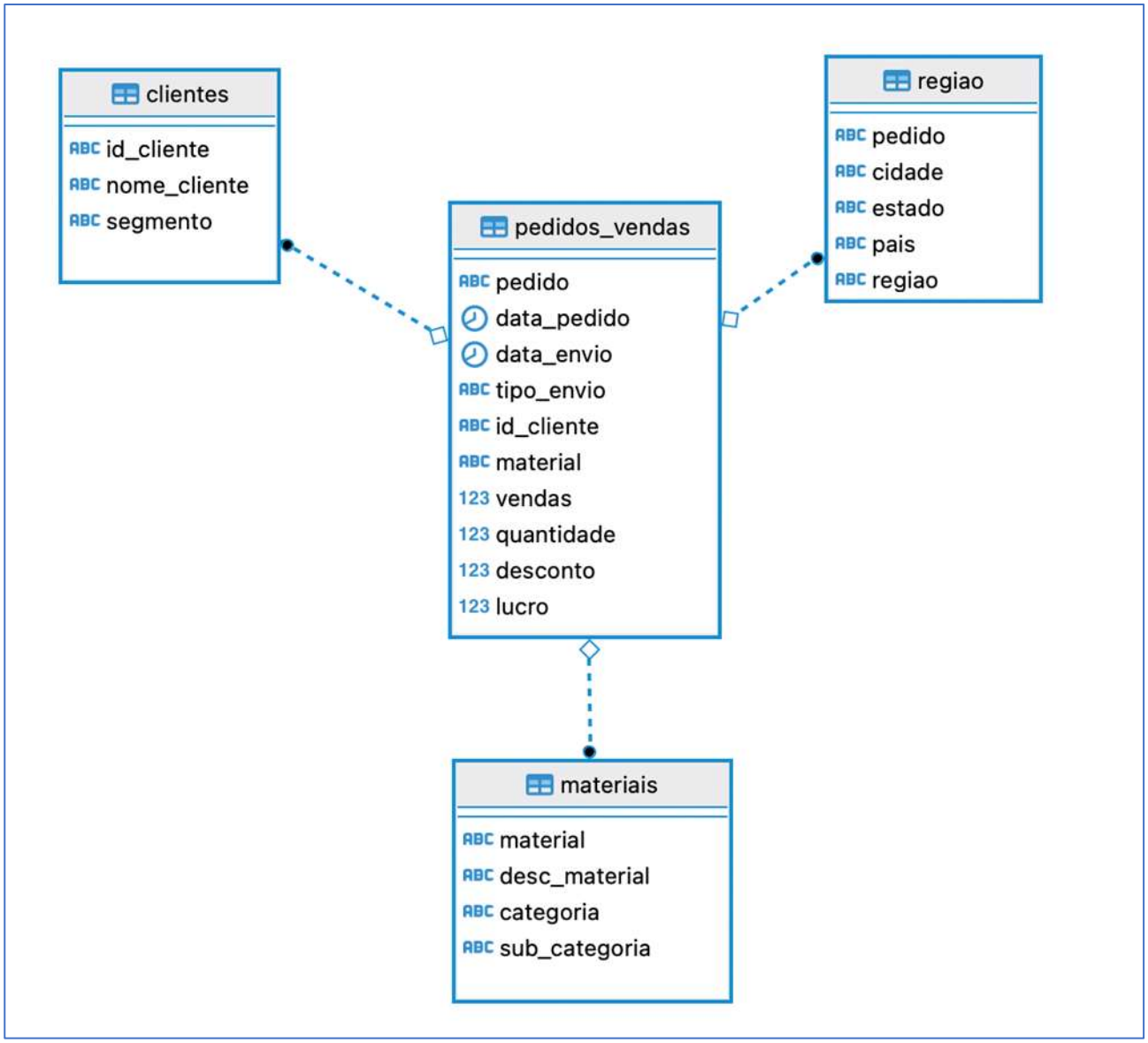
Campo	Tipo de Dados	Descrição
pedido	STRING	Pedidos realizados
data_pedido	DATA	Data de criação do pedido
data_envio	DATA	Data de envio
tipo_envio	STRING	Modalidade de envio do material
nome_cliente	STRING	Nome do cliente que fez a compra
segmento	STRING	Segmento de atuação do cliente
material	STRING	Código único do material
desc_material	STRING	Descrição completa do material
categoria	STRING	Categoria que o material pertence
sub_categoria	STRING	Sub-Categoria do material
cidade	STRING	Cidade de envio
regiao	STRING	Região
pais	STRING	País
vendas	FLOAT	Total de vendas realizadas em $
quantidade	FLOAT	Quantidade de produtos vendidos
desconto	FLOAT	Desconto aplicado
lucro	FLOAT	Lucro obtido
upd_timestamp	TIMESTAMP	Essa coluna será criada apenas nas tabelas CLEAN como controle de versão e atualização

4 - Modelagem de Dados

Tendo como base o dataset superloja, vamos tratar o conceito RAW (tabela bruta) e tabela CLEAN (dados limpos).

- Tabela RAW: São os dados importados diretamente da origem (seja ela qual for) direto para o Data Warehouse Cloud.
- Tabela CLEN: São os dados tratados, limpos e mais atualizados de forma distinta.

Modelo de dados Proposto (Conceitual)



4.1 - Modelagem de Dados | Scripts de Criação

Neste item, foram criadas as tabelas CLEAN no Data Wharehouse. Para tal, foram utilizadas as Plataformas Google Cloud e o BigQuery como Data Warehouse.

Tabelas CLEAN

```
--- Tabela Pedidos
CREATE TABLE `mydatasets_clean.pedidos_vendas` (
    pedido STRING NULL,
    data_pedido date NULL,
    data_envio date NULL,
    tipo_envio STRING NULL,
    id_cliente STRING NULL,
    material STRING NULL,
    vendas FLOAT NULL,
    quantidade INTEGER NULL,
    desconto FLOAT NULL,
    lucro FLOAT NULL,
    upd_timestamp TIMESTAMP NULL
);

--- Tabela Material
CREATE TABLE `mydatasets_clean.materiais` (
    material STRING NULL,
    desc_material STRING NULL,
    categoria STRING NULL,
    sub_categoria STRING NULL,
    upd_timestamp TIMESTAMP NULL
);

--- Tabela Região
CREATE TABLE `mydatasets_clean.regiao` (
    pedido STRING NULL,
    cidade STRING NULL,
    estado STRING NULL,
    pais STRING NULL,
    regiao STRING NULL,
    upd_timestamp TIMESTAMP NULL
);

--- Tabela Clientes
CREATE TABLE `mydatasets_clean.clientes` (
    id_cliente STRING NULL,
    nome_cliente STRING NULL,
    segmento STRING NULL,
    upd_timestamp TIMESTAMP NULL
);
```

A Tabela RAW está localizada em `mydatasets.superloja_raw` e foi criada a partir de um Google Sheets.

5 - Carga

Segue abaixo os scripts de atualização de dados a serem configurados para execução diária (D-1) ou a cada hora dependendo do ambiente. Foi adotada a atualização D-1.

```
-----------------------------------------------------
-- TABELA pedidos_vendas
-----------------------------------------------------
TRUNCATE TABLE `mydatasets_clean.pedidos_vendas`;

INSERT INTO `mydatasets_clean.pedidos_vendas`
SELECT DISTINCT
  pedido,
  data_pedido,
  data_envio,
  tipo_envio,
  id_cliente,
  material,
  vendas,
  quantidade,
  desconto,
  lucro,
  current_timestamp() AS upd_timestamp
FROM `mydatasets.superloja_raw`
;

-----------------------------------------------------
-- TABELA materiais
-----------------------------------------------------
TRUNCATE TABLE `mydatasets_clean.materiais`;

INSERT INTO `mydatasets_clean.materiais`
SELECT DISTINCT
  material,
  desc_material,
  categoria,
  sub_categoria,
  current_timestamp() AS upd_timestamp
FROM `mydatasets.superloja_raw`
;

-----------------------------------------------------
-- TABELA clientes
-----------------------------------------------------
TRUNCATE TABLE `mydatasets_clean.clientes`;

INSERT INTO `mydatasets_clean.clientes`
SELECT DISTINCT
  id_cliente,
  nome_cliente,
  segmento,
    current_timestamp() AS upd_timestamp
FROM `mydatasets.superloja_raw`
;


-----------------------------------------------------
-- TABELA regiao
-----------------------------------------------------
TRUNCATE TABLE `mydatasets_clean.regiao`;

INSERT INTO `mydatasets_clean.regiao`
SELECT DISTINCT
  pedido,
```

```
  cidade,
  estado,
  pais,
  regiao,
   current_timestamp() AS upd_timestamp
FROM `mydatasets.superloja_raw`
;
```

5.1 - Agendamento da Atualização ETL

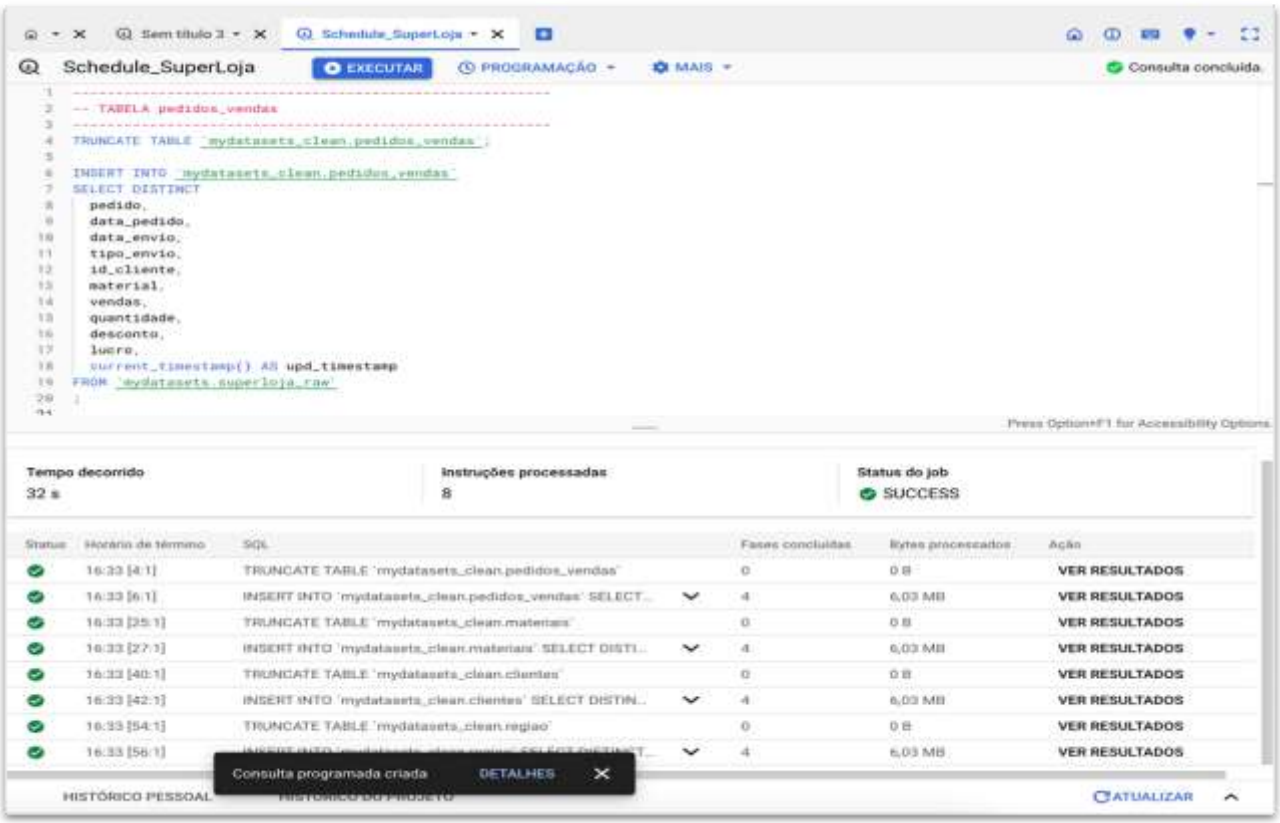
Abaixo, a criação do agendamento do processo de carga dos dados. A partir dos scripts descritos acima, foi agendado o processo automático de carga.

Para este fim, foi utilizado a ferramenta Schedule (ou programação de Consultas) no BigQuery. No Schedule, podemos programar várias consultas para serem executadas a qualquer momento, seja a cada hora ou por dia, por mês, etc.

Neste caso, foi desenvolvido um schedule para ser executado uma vez ao dia, onde irá atualizar diariamente as tabelas CLEAN.

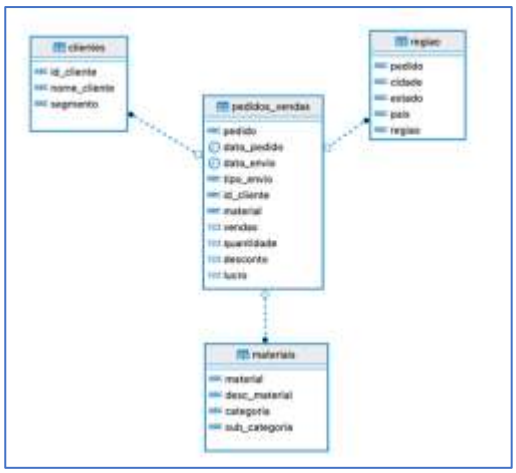
Abaixo, temos schedule criado e programado:

Nesta imagem, temos o conteúdo do Schedule:

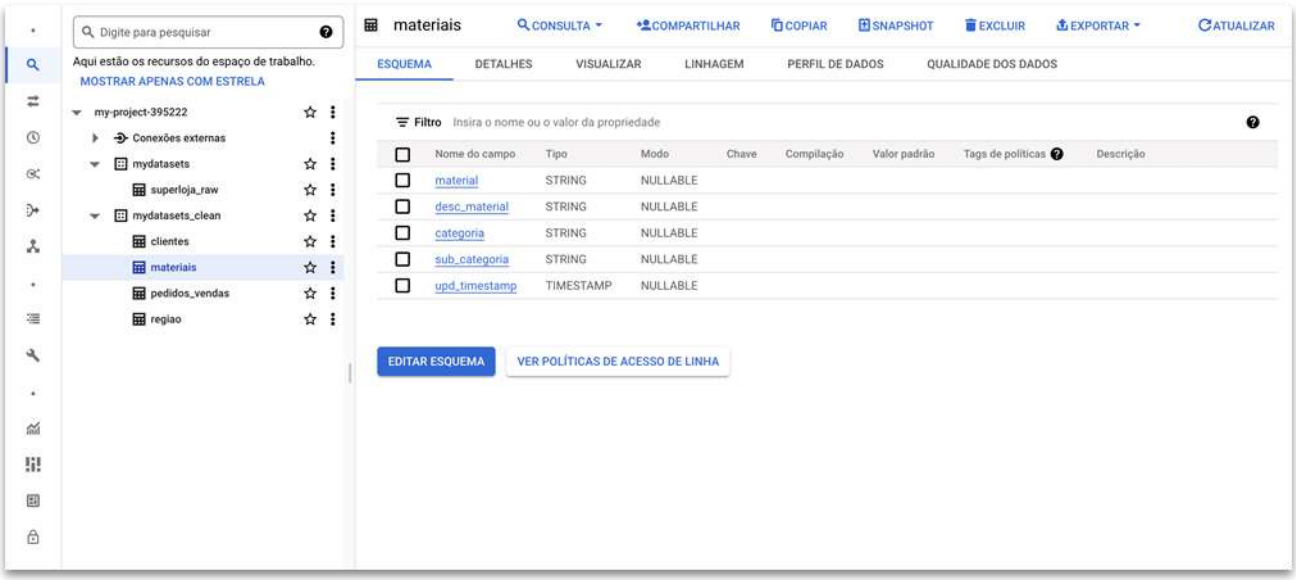


6 - Análise de Dados | Qualidade de Dados

No dataset da Super Loja, não foram encontrados nenhum problema adicional, porém para melhor performance e ter uma base de dados equalizada, foi criada a modelagem "Esquema Estrela" .



Abaixo segue as tabelas criadas no BigQuery:



7 - Análise de Dados | Solução do Problema

Abaixo, as respostas para as perguntas de negócio em análise para a Solução do Problema relatado nos Objetivos. Para tal, foram desenvolvidas consultas SQL no BigQuery.

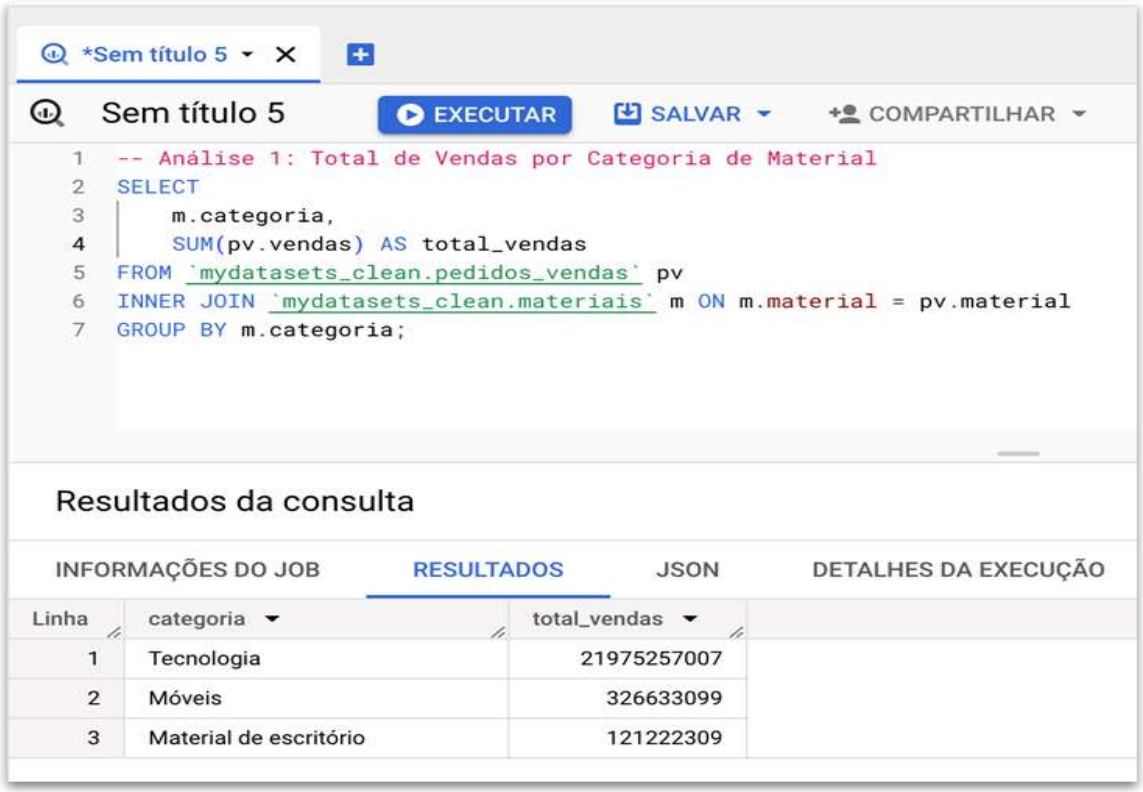
Perguntas a serem respondidas são:

a) Análise 1: Qual o Total de Vendas por Categoria de Material
b) Análise 2: Mostre o Lucro Total por Região
c) Análise 3: Qual a Média de Desconto por Segmento de Clientes
d) Análise 4: Faça uma Consulta que mostre as Vendas por Mês
e) Análise 5: Qual o Lucro por Material
f) Análise 6: Qual é a Quantidade de Pedidos por País
g) Análise 7: Faça uma consulta que mostre o Top 10 Clientes com Maior Valor de Vendas
h) Análise 8: Mostre o Total de Vendas por Ano e Segmento de Clientes

7.1 - Scripts

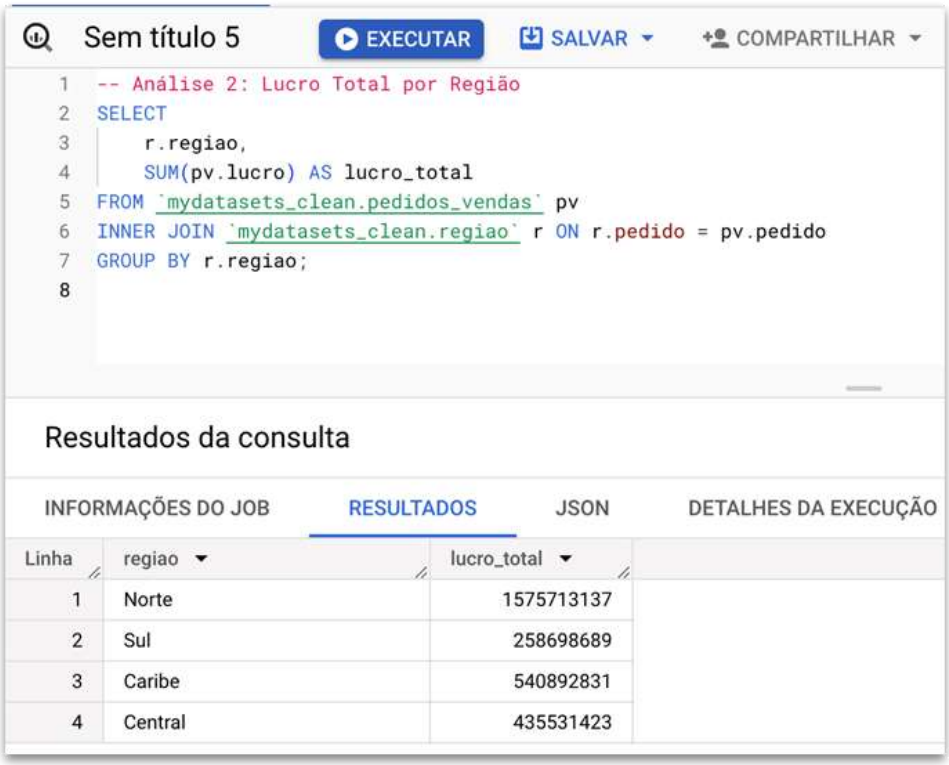
a) Análise 1: Qual o Total de Vendas por Categoria de Material

```
-- Análise 1: Total de Vendas por Categoria de Material
SELECT
    m.categoria,
    SUM(pv.vendas) AS total_vendas
FROM `mydatasets_clean.pedidos_vendas` pv
INNER JOIN `mydatasets_clean.materiais` m ON m.material = pv.material
GROUP BY m.categoria;
```



b) Análise 2: Mostre o Lucro Total por Região

```
-- Análise 2: Lucro Total por Região
SELECT
    r.regiao,
    SUM(pv.lucro) AS lucro_total
FROM `mydatasets_clean.pedidos_vendas` pv
INNER JOIN `mydatasets_clean.regiao` r ON r.pedido = pv.pedido
GROUP BY r.regiao;
```



c) Análise 3: Qual a Média de Desconto por Segmento de Clientes

```
-- Análise 3: Média de Desconto por Segmento de Clientes
SELECT
    c.segmento,
    AVG(pv.desconto) AS media_desconto
FROM `mydatasets_clean.pedidos_vendas` pv
INNER JOIN `mydatasets_clean.clientes` c ON c.id_cliente = pv.id_cliente
GROUP BY c.segmento
ORDER BY 2 DESC;
```

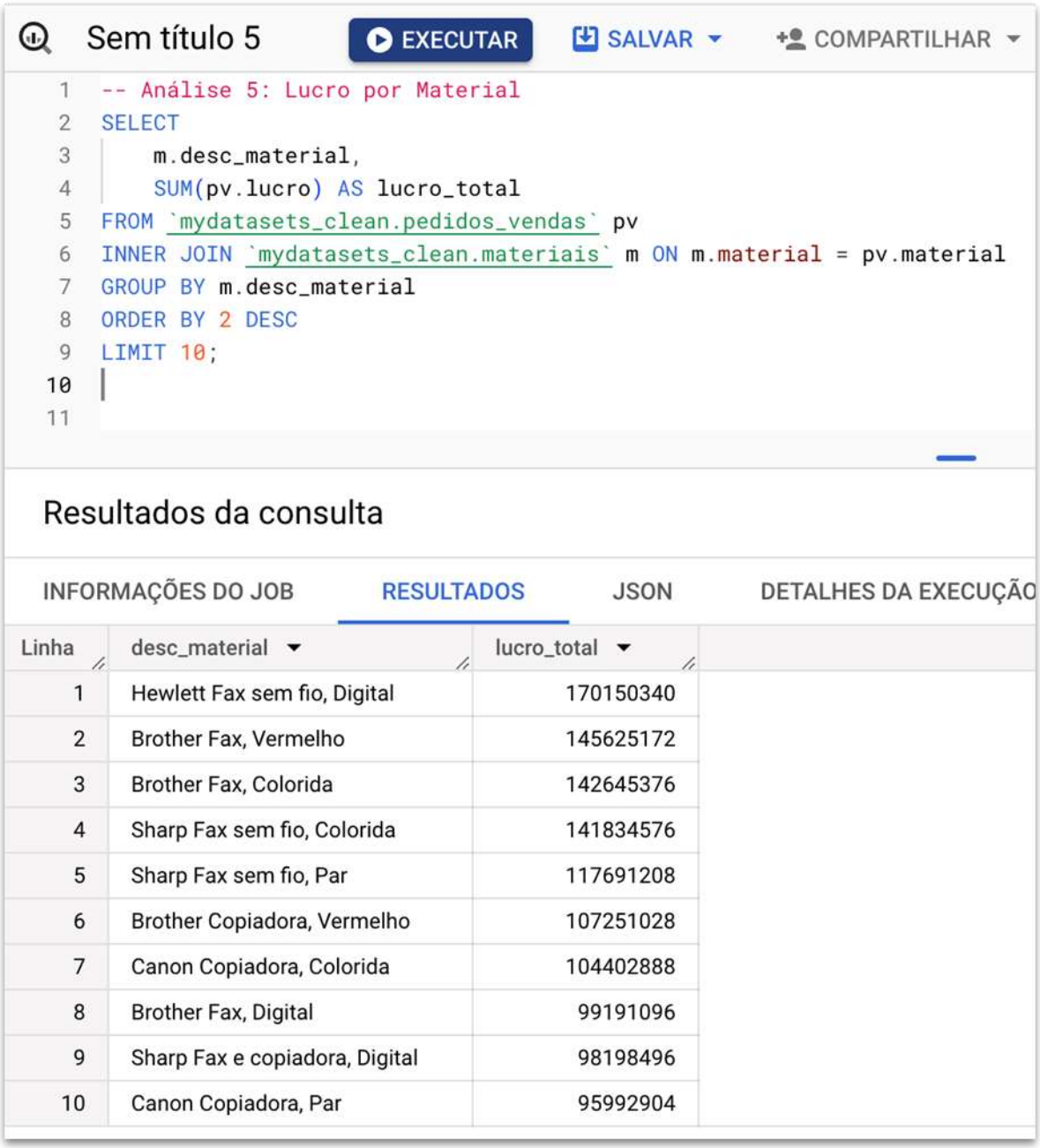
d) Análise 4: Faça uma Consulta que mostre as Vendas por Mês

```
-- Análise 4: Vendas por Mês
SELECT
    DATE_FORMAT(pv.data_pedido, '%Y-%m') AS mes,
    SUM(pv.vendas) AS total_vendas
FROM `mydatasets_clean.pedidos_vendas` pv
GROUP BY mes
ORDER BY 1 DESC;
```

Linha	mes	total_vendas
1	2023-12	627355814
2	2023-11	1183859550
3	2023-10	1006913737
4	2023-09	1133148494
5	2023-08	418604401
6	2023-07	603687730
7	2023-06	767643899
8	2023-05	324079366
9	2023-04	376158987
10	2023-03	291345261
11	2023-02	217808709
12	2023-01	124778352
13	2022-12	480326470
14	2022-11	621825979
15	2022-10	726093262
16	2022-09	413820623
17	2022-08	356724268

e) Análise 5: Qual o Lucro por Material

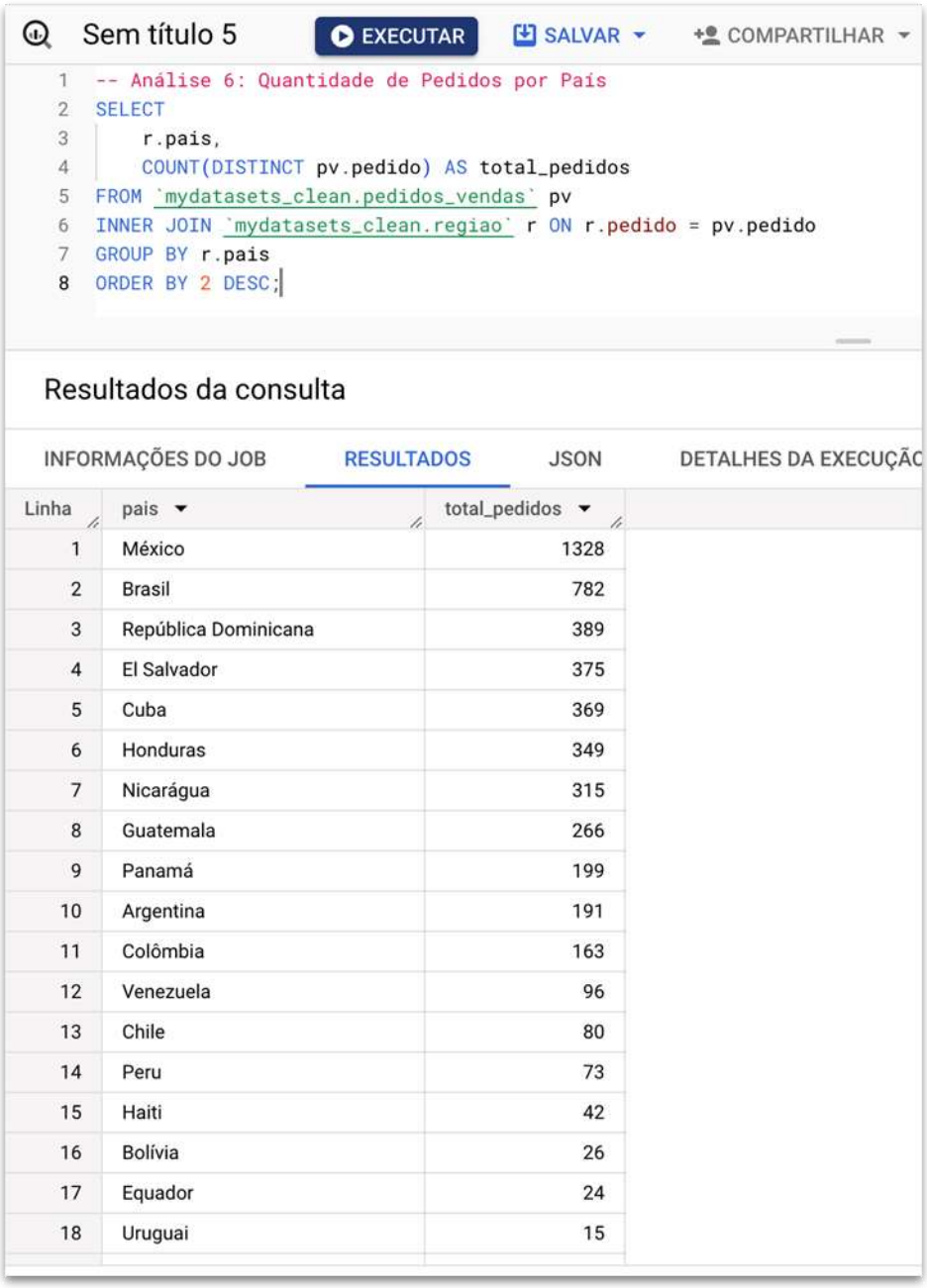
```
-- Análise 5: Lucro por Material
SELECT
    m.desc_material,
    SUM(pv.lucro) AS lucro_total
FROM `mydatasets_clean.pedidos_vendas` pv
INNER JOIN `mydatasets_clean.materiais` m ON m.material = pv.material
GROUP BY m.desc_material
ORDER BY 2 DESC
LIMIT 10;
```



Linha	desc_material	lucro_total
1	Hewlett Fax sem fio, Digital	170150340
2	Brother Fax, Vermelho	145625172
3	Brother Fax, Colorida	142645376
4	Sharp Fax sem fio, Colorida	141834576
5	Sharp Fax sem fio, Par	117691208
6	Brother Copiadora, Vermelho	107251028
7	Canon Copiadora, Colorida	104402888
8	Brother Fax, Digital	99191096
9	Sharp Fax e copiadora, Digital	98198496
10	Canon Copiadora, Par	95992904

f) Análise 6: Qual é a Quantidade de Pedidos por País

```
-- Análise 6: Quantidade de Pedidos por País
SELECT
    r.pais,
    COUNT(DISTINCT pv.pedido) AS total_pedidos
FROM `mydatasets_clean.pedidos_vendas` pv
INNER JOIN `mydatasets_clean.regiao`  r ON r.pedido = pv.pedido
GROUP BY r.pais
ORDER BY 2 DESC;
```



Linha	pais	total_pedidos
1	México	1328
2	Brasil	782
3	República Dominicana	389
4	El Salvador	375
5	Cuba	369
6	Honduras	349
7	Nicarágua	315
8	Guatemala	266
9	Panamá	199
10	Argentina	191
11	Colômbia	163
12	Venezuela	96
13	Chile	80
14	Peru	73
15	Haiti	42
16	Bolívia	26
17	Equador	24
18	Uruguai	15

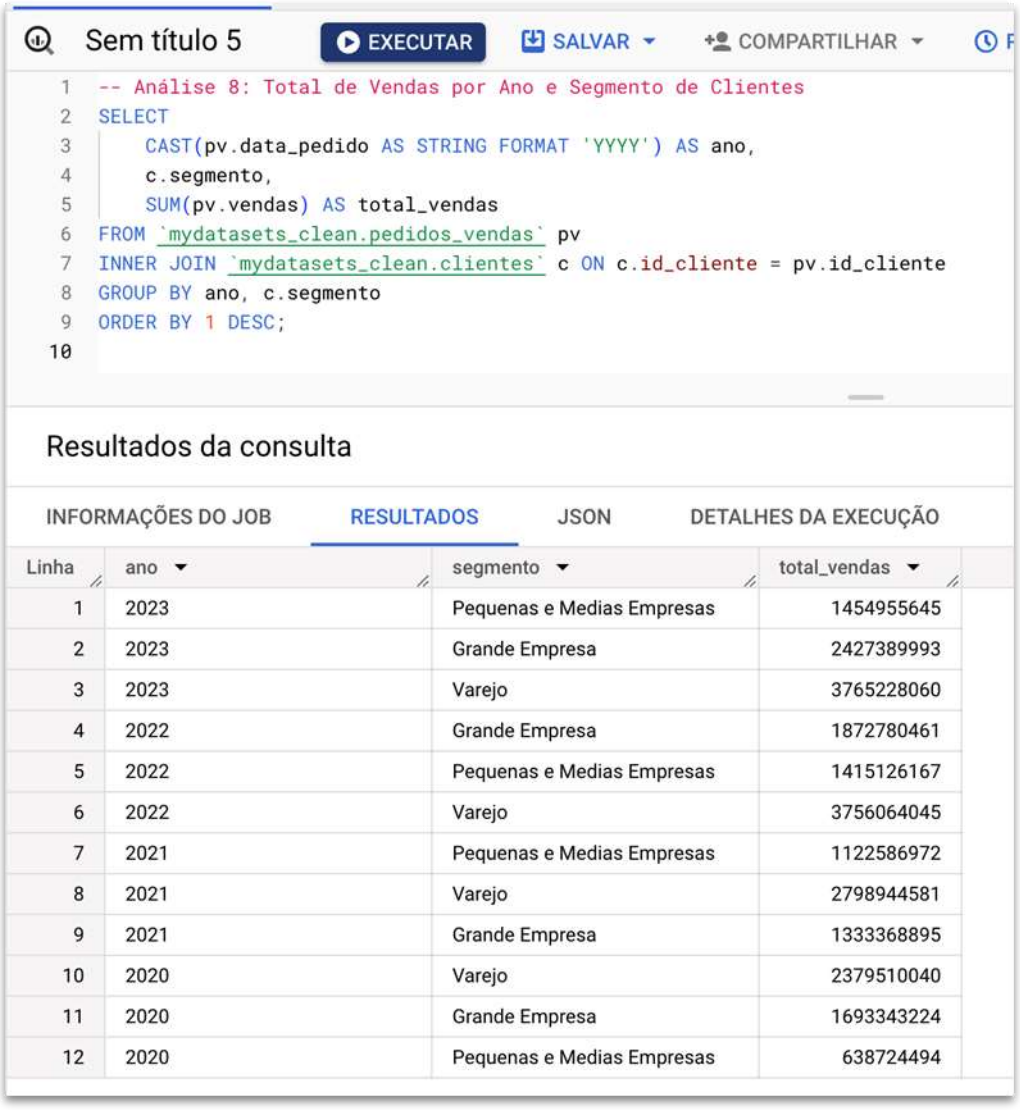
g) Análise 7: Faça uma consulta que mostre o Top 10 Clientes com Maior Valor de Vendas

```
-- Análise 7: Top 10 Clientes com Maior Valor de Vendas
SELECT
    c.nome_cliente,
    SUM(pv.vendas) AS total_vendas
FROM `mydatasets_clean.pedidos_vendas` pv
INNER JOIN `mydatasets_clean.clientes` c ON c.id_cliente = pv.id_cliente
GROUP BY c.nome_cliente
ORDER BY total_vendas DESC
LIMIT 10;
```

Linha	nome_cliente	total_vendas
1	Vinicius Silva	550365850
2	Beatriz Goncalves	428585270
3	Davi Ferreira	387365914
4	Marisa Araujo	322593374
5	Leonardo Rodrigues	311898594
6	Paulo Alves	205328348
7	Alex Dias	205146064
8	Murilo Carvalho	198331899
9	Marisa Almeida	197609246
10	Enzo Azevedo	196682538

h) Análise 8: Mostre o Total de Vendas por Ano e Segmento de Clientes

```
-- Análise 8: Total de Vendas por Ano e Segmento de Clientes
SELECT
    CAST(pv.data_pedido AS STRING FORMAT 'YYYY') AS ano,
    c.segmento,
    SUM(pv.vendas) AS total_vendas
FROM `mydatasets_clean.pedidos_vendas` pv
INNER JOIN `mydatasets_clean.clientes` c ON c.id_cliente = pv.id_cliente
GROUP BY ano, c.segmento
ORDER BY 1 DESC;
```



Linha	ano	segmento	total_vendas
1	2023	Pequenas e Medias Empresas	1454955645
2	2023	Grande Empresa	2427389993
3	2023	Varejo	3765228060
4	2022	Grande Empresa	1872780461
5	2022	Pequenas e Medias Empresas	1415126167
6	2022	Varejo	3756064045
7	2021	Pequenas e Medias Empresas	1122586972
8	2021	Varejo	2798944581
9	2021	Grande Empresa	1333368895
10	2020	Varejo	2379510040
11	2020	Grande Empresa	1693343224
12	2020	Pequenas e Medias Empresas	638724494

8 - Conclusão | Autoavaliação

Durante o processo de desenvolvimento do projeto, foi evidenciado que o conjunto de dados selecionado já se encontrava devidamente formatado e tratado, exigindo apenas ajustes mínimos, com exceção da necessidade de implementação de um modelo de dados em formato Estrela.

Contudo, cabe lembrar um dos pontos amplamente discutido nas aulas, que "na vida real" nem sempre os dados não estarão perfeitos e sem erros como no caso de estudo do dataset da "SuperLoja", particularmente no que se refere ao cadastro, abrangendo fornecedores, clientes, materiais, entre outros. Tais erros podem ter impactos significativos na operação do Data Warehouse, dependendo da gravidade dos mesmos, podendo até mesmo interromper parcialmente suas funcionalidades.

Por essa razão, constatei a importância do investimento em arquitetura de dados, bem como na criação de um catálogo de dados, Governança de dados, dentre outras iniciativas. Cabe ressaltar que alguns desses problemas podem ser mitigados durante o processo ETL, embora nem todos possam ser completamente corrigidos.

Ficou claro para mim, durante a elaboração do presente MVP, a importância da sintonia e sinergia entre as áreas técnicas e de negócio, para promover a análise e formulação das "perguntas" e questões de negócios a serem respondidas através dos dados internos e externos das organizações. A minha experiência na área comercial, ajudou, sobremaneira, escolhas das "perguntas de negócios" e respectiva avaliação dos datasets para respondê-las. Tal constatação, reforça a importância da colaboração estreita entre as equipes de TI e Negócios, uma vez que ambas são interdependentes.

Meu grande desafio foi executar, na prática, as etapas de ETL, principalmente utilizando ambientes e ferramentas em nuvem. Depois de muitas tentativas, erros e pesquisas de soluções, consegui atingir os resultados acima alcançados. Certamente irei aumentar o meu respeito e admiração e importância da atuação dos profissionais de tratamento dos dados, pois são "engrenagem" fundamental dos projetos e soluções de Data & Analytcs.

Destaco, por fim, minha apreciação pelo uso das ferramentas disponibilizadas pelo GCP, especialmente o BigQuery, que facilitou na elaboração e conclusão do trabalho. De qualquer forma, continuarei o teste e uso de outras plataformas como AWS e Azure nos futuros projetos.