

Objetivos do Projeto

O projeto tem como objetivo entender o processo de criação de um banco de dados para uso analítico enquanto utiliza temas comuns no mercado de trabalho, como o conhecimento das camadas de um banco de dados, as regras e métricas de negócio envolvidas em uma plataforma digital de vendas em áreas como finanças, marketing e vendas, o uso de plataformas bastante populares de tratamento de dados, como o databricks e das linguagens mais utilizadas, como o SQL.

A ideia, à princípio, é encontrar uma base de dados que disponibilize dados relativos a vendas de um negócio. Plataformas como o Kaggle e fóruns como o Reddit costumam disponibilizar bases públicas de boa usabilidade e que permitem realizar análises relevantes e muito semelhantes a cases reais do varejo.

O projeto visa implementar um datalake com arquitetura em três camadas (bronze, silver e gold) que possibilite armazenar e organizar dados transacionais de vendas em sua camada bronze, transformar dados brutos em formatos limpos e padronizados em sua camada silver e, por último, criar visualizações úteis para análises de negócio em sua camada gold. O projeto deverá ser documentado e executado conforme os padrões de boas práticas da área.

Dentre as perguntas que devem ser respondidas pelas visualizações, estão:

-
- Quais produtos têm melhor desempenho de vendas?
- Como as vendas variam ao longo do tempo?
- Qual é a eficácia de diferentes campanhas promocionais?
- Qual é o desempenho comparativo entre diferentes lojas?
- Qual é o desempenho comparativo entre diferentes regiões?
- Quais produtos e categorias geram maiores margens de lucro?
- Podemos manter um estoque dinâmico de produtos com base na sazonalidade do negócio?
- Quais são os padrões de comportamento dos clientes?
- Qual o perfil dos nossos clientes em relação a idade, sexo, regiões do país?

A resposta dessas perguntas permitirá ao negócio a otimização de estoque e campanhas de marketing, além de demonstrar a necessidade de lojas em determinado país ou cidade e o perfil do cliente nesses locais. O objetivo de uma análise, em geral, é otimizar o negócio visando o aumento de performance, seja local ou globalmente.

Etapas 1 e 2 - Busca dos dados e Coleta

Para encontrar dados que se encaixassem ao objetivo principal do projeto, recorri ao Kaggle - ferramenta que já tenho o hábito de utilizar e sei que possui uma variedade imensa de datasets, dos quais ao menos um se encaixaria em meus objetivos. Testados alguns datasets, o escolhido foi o [Global Fashion Retail Sales](#) pela estrutura dos seus dados, volume, usabilidade e quantidade de tabelas disponibilizadas - grande o suficiente para permitir demonstrar o conhecimento em joins, criação de entidades e relacionamentos e análise dos dados. Nessa etapa o dataset [Chocolate Sales Data](#) também foi cogitado, mas não foi o escolhido por conta da baixa quantidade de tabelas e colunas disponíveis, que tornaria o trabalho como um todo menos proveitoso para o aprendizado. Os arquivos foram baixados em CSV e foi feito o upload no Databricks, plataforma utilizada durante as aulas, conforme demonstrado em tópicos mais à frente.

Etapa 3 - Modelagem dos Dados

Para o banco de dados, foi escolhida uma modelagem em esquema flat pelas características do negócio e dos dados escolhidos. O esquema flat, embora consuma mais espaço de armazenamento, são mais eficazes e rápidos em consultas complexas, o que não é verdade em um star schema, dado que as quantidade de junções entre fatos e dimensões exigem um tempo considerável de processamento. O star schema também é pouco flexível quanto a mudanças na estrutura das tabelas ou eventual adição de informações, o que torna o processo muito trabalhoso quando comparado a um modelo flat.

O banco de dados também foi dividido em três camadas: bronze, silver e gold. Na camada bronze, os dados são ingeridos em seu formato bruto para que sejam preservadas as características originais dos dados e, se necessário por conta de algum erro operacional, possa existir a reconstituição das camadas seguintes, garantindo que nenhuma informação será perdida durante o processo de tratamento e consumo dos dados e mantendo um histórico verdadeiro. A camada silver é responsável por armazenar pequenas alterações nos dados, como o uso de comandos que limpam possíveis espaços vazios no início e fim das strings, tratamento de datas e timestamps, renomear colunas (que, no caso do dataset utilizado, possuía espaços e caracteres especiais, o que não é uma boa prática). Pode ser considerada a camada intermediária entre a bronze e a gold: nem tão “crua” quanto a primeira, nem tão

refinada quanto a segunda. Já a camada gold é responsável por trazer os dados de forma que as análises de negócio sejam possíveis de maneira rápida, o que implica na realização de cálculos, agregações, agrupamentos e criação de novas colunas com informações enriquecidas. O conjunto de dados escolhido conta com 6 tabelas: customers, referente aos dados dos clientes; discounts, referentes aos tipos de descontos disponíveis e suas datas; employees, referente aos dados dos funcionários responsáveis pela venda; products, referente aos dados dos produtos disponíveis em loja; stores, referente aos dados das lojas existentes e transactions, referente aos dados das transações financeiras feitas pelos clientes.

A estrutura dos dados se dá conforme o dicionário a seguir - os valores possíveis foram encontrados através de consultas utilizando *select distinct* diretamente nas tabelas para variáveis categóricas e *select min(data) from tabela/select max(data) from tabela* para variáveis de data e numéricos, e estão adequados ao formato que os dados têm nas tabelas:

1. Camada Silver:

Tabela Customers:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
customer_id	STRING	Identificador único do cliente	Alfanumérico	-
customer_name	STRING	Nome completo do cliente	Texto	-
email	STRING	Endereço de email do cliente	Email	-
telephone	STRING	Número de telefone do cliente	Alfanumérico	-
city	STRING	Cidade de residência do cliente	Texto	-
country	STRING	País de residência do cliente	Texto	Textos de 3 dígitos
gender	STRING	Gênero do cliente	Categórico	F, M, D
date_of_birth	DATE	Data de nascimento	Data	1940-01-01 a 2010-12-31

		cliente		
job_title	STRING	Título do cargo/profissão do cliente	Texto	-

Tabela Discounts:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
start_date	DATE	Data de início da campanha de desconto	Data	2000-01-01 a 2030-12-31
end_date	DATE	Data de término da campanha de desconto	Data	2000-01-01 a 2030-12-31
discount_percentag e	DECIMAL(5,2)	Percentual de desconto oferecido	Numérico	1.00 a 99.00
description	STRING	Descrição da campanha de desconto	Texto	-
category	STRING	Categoria de produto aplicável ao desconto	Categórico	"Clothing", "Accessories", "Footwear", etc.
sub_category	STRING	Subcategoria de produto aplicável ao desconto	Categórico	"Shirts", "Dresses", "Pants", "Handbags", etc.

Tabela Employees:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
employee_id	STRING	Identificador único do funcionário	Alfanumérico	-
store_id	STRING	Identificador da loja onde o funcionário trabalha	Alfanumérico	-
employee_name	STRING	Nome completo do funcionário	Texto	-

		funcionário		
position	STRING	Cargo do funcionário na empresa	Categórico	"Sales Associate", "Store Manager", "Cashier", "Stock Clerk", "Assistant Manager"

Tabela Products:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
product_id	STRING	Identificador único do produto	Alfanumérico	-
category	STRING	Categoria principal do produto	Categórico	"Masculine", "Feminine", "Children".
sub_category	STRING	Subcategoria do produto	Categórico	"Lingerie and Pajamas", "Suits and Blazers", "Coats and Blazers", "Coats", "Girl and Boy (1-5 years, 6-14 years)", "Skirts and Shorts", "T-shirts and Tops", "Underwear and Pajamas", "Sweaters and Knitwear", "Sportswear", "Pants and Jeans", "Dresses and Jumpsuits", "Sweaters and Sweatshirts", "Pajamas",

				"Shirts and Blouses", "Suits and Sets", "Accessories", "Shirts", "T-shirts and Polos", "Baby (0-12 months)", "Sweaters"
description_pt	STRING	Descrição do produto em português	Texto	-
description_de	STRING	Descrição do produto em alemão	Texto	-
description_fr	STRING	Descrição do produto em francês	Texto	-
description_es	STRING	Descrição do produto em espanhol	Texto	-
description_en	STRING	Descrição do produto em inglês	Texto	-
description_zh	STRING	Descrição do produto em chinês	Texto	-
color	STRING	Cor(es) disponível(is) do produto	Categórico	"RED", "LILAC", "WHITE", "BLACK", "BEIGE", null, "GREEN", "SILVER", "GOLD", "YELLOW",

				“NEUTRAL”, “TURQUOISE”, “BLUE”, “PINK”, “BURGUNDY”, “MUSTARD”
sizes	STRING	Tamanhos disponíveis do produto	Categórico	“P M G GG”, null “38 40 42 44 46 48” , “36 38 40 42 44 46” , “M L XL”, “38 40 42 44 46”, “38 40 42 44”, “S M L XL”, “36 38 40 42 44”, “M L XL XXL”, “S M L”, “P M G”, “36 38 40 42”
production_cost	DECIMAL(10,2)	Custo de produção do produto	Numérico	0.01 a 50000.00

Tabela Stores:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
store_id	STRING	Identificador único da loja	Alfanumérico	-
country	STRING	País onde a loja está localizada	Categórico	“España”, “France”, “United States”, “中国”, “Deutschland”, “Portugal”, “United Kingdom”

city	STRING	Cidade onde a loja está localizada	Texto	-
store_name	STRING	Nome comercial da loja	Texto	-
number_of_employees	INT	Número de funcionários da loja	Numérico	1 a 50000
zip_code	STRING	Código postal da loja	Alfanumérico	-
latitude	DOUBLE	Coordenada geográfica - latitude	Numérico	-90.0 a 90.0
longitude	DOUBLE	Coordenada geográfica longitude	Numérico	-180.0 a 180.0

Tabela Transactions:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
invoice_id	STRING	Identificador único da fatura	Alfanumérico	-
line_number	INT	Número sequencial do item na fatura	Numérico	1 a 1000
customer_id	STRING	Identificador do cliente que realizou a compra	Alfanumérico	-
product_id	STRING	Identificador do produto comprado	Alfanumérico	-
size	STRING	Tamanho do produto comprado	Categórico	-
color	STRING	Cor do produto comprado	Categórico	-
unit_price	DECIMAL(10,2)	Preço unitário do produto	Numérico	1.00 a 100000.00
quantity	INT	Quantidade de	Numérico	1 a 50

		itens comprados		
transaction_date	DATE	Data da transação	Data	2000-01-01 a 2030-12-31
discount_percentag e	DECIMAL(5,2)	Percentual de desconto aplicado	Numérico	0.01 a 99.00
line_total	DECIMAL(12,2)	Valor total da linha (quantidade * preço com desconto)	Numérico	5.00 a 10000.00
store_id	STRING	Identificador da loja onde ocorreu a venda	Alfanumérico	-
employee_id	STRING	Identificador do funcionário que realizou a venda	Alfanumérico	-
currency	STRING	Moeda utilizada na transação	Categórico	"USD", "EUR", "BRL", "CNY", etc.
currency_symbol	STRING	Símbolo da moeda utilizada	Categórico	"\$", "€", "R\$", "¥", etc.
sku	STRING	Código de referência do produto (Stock Keeping Unit)	Alfanumérico	-
transaction_type	STRING	Tipo de transação	Categórico	"Purchase", "Return", "Exchange"
payment_method	STRING	Método de pagamento utilizado	Categórico	"Credit Card", "Debit Card", "Cash", "PayPal", etc.
invoice_total	DECIMAL(12,2)	Valor total da fatura	Numérico	5.00 a 50000.00

2. Camada Gold:

Tabela DAILY_SALES_BY_CATEGORY:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
transaction_date	DATE	Data da transação	Data	2020-01-01 a 2023-12-31
category	STRING	Categoria principal do produto	Categórico	"Clothing", "Accessories", "Footwear", etc.
sub_category	STRING	Subcategoria do produto	Categórico	"Shirts", "Dresses", "Pants", "Handbags", etc.
num_transactions	BIGINT	Número de transações distintas	Numérico	0 a 10000
total_items_sold	BIGINT	Total de itens vendidos	Numérico	0 a 100000
total_revenue	DECIMAL(16,2)	Receita total	Numérico	0.00 a 1000000.00
total_profit	DECIMAL(16,2)	Lucro total (receita - custo)	Numérico	-10000.00 a 500000.00
avg_discount	DECIMAL(5,2)	Desconto médio aplicado	Numérico	0.00 a 70.00

Tabela CUSTOMER_PROFILE:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
customer_id	STRING	Identificador único do cliente	Alfanumérico	-
customer_name	STRING	Nome completo do cliente	Texto	-
gender	STRING	Gênero do cliente	Categórico	"Male", "Female", "Non-binary", "Prefer not to say"
country	STRING	País de residência do cliente	Texto	"Brazil", "USA", "France", "Germany", etc.

city	STRING	Cidade de residência do cliente	Texto	Conjunto de cidades diversas
job_title	STRING	Título do cargo/profissão do cliente	Texto	"Engineer", "Teacher", "Manager", etc.
age	INT	Idade calculada do cliente	Numérico	18 a 85
total_purchases	BIGINT	Número total de compras realizadas	Numérico	0 a 1000
total_spent	DECIMAL(16,2)	Valor total gasto pelo cliente	Numérico	0.00 a 100000.00
avg_purchase_value	DECIMAL(16,2)	Valor médio por compra	Numérico	0.00 a 10000.00
last_purchase_date	DATE	Data da última compra	Data	2020-01-01 a 2023-12-31
days_since_last_purchase	INT	Número de dias desde a última compra	Numérico	0 a 1000
preferred_categories	ARRAY<STRING>	Categorias preferidas (mais compradas)	Array	["Clothing", "Accessories"]

Tabela STORE_PERFORMANCE:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
store_id	STRING	Identificador único da loja	Alfanumérico	-
store_name	STRING	Nome comercial da loja	Texto	"Fashion Store Downtown", "Fashion Mall Plaza", etc.
country	STRING	País onde a loja está localizada	Categórico	"Brazil", "USA", "France",

				"Germany", etc.
city	STRING	Cidade onde a loja está localizada	Texto	Conjunto de cidades diversas
number_of_employees	INT	Número de funcionários da loja	Numérico	3 a 50
total_transactions	BIGINT	Número total de transações	Numérico	0 a 100000
total_revenue	DECIMAL(16,2)	Receita total da loja	Numérico	0.00 a 10000000.00
revenue_per_employee	DECIMAL(16,2)	Receita por funcionário	Numérico	0.00 a 1000000.00
unique_customers	BIGINT	Número de clientes únicos	Numérico	0 a 50000
avg_discount	DECIMAL(5,2)	Desconto médio aplicado	Numérico	0.00 a 70.00
total_profit	DECIMAL(16,2)	Lucro total da loja	Numérico	-1000000.00 a 5000000.00
profit_margin	DECIMAL(6,2)	Margem de lucro percentual	Numérico	-20.00 a 80.00

Tabela EMPLOYEE_PERFORMANCE:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
employee_id	STRING	Identificador único do funcionário	Alfanumérico	-
employee_name	STRING	Nome completo do funcionário	Texto	-
position	STRING	Cargo do funcionário na empresa	Categórico	"Sales Associate", "Store Manager", "Cashier", etc.
store_id	STRING	Identificador da loja onde o funcionário trabalha	Alfanumérico	-
store_name	STRING	Nome comercial da	Texto	"Fashion Store

		loja		Downtown", "Fashion Mall Plaza", etc.
total_sales	BIGINT	Número total de vendas realizadas	Numérico	0 a 10000
total_revenue	DECIMAL(16,2)	Receita total gerada	Numérico	0.00 a 1000000.00
avg_sale_value	DECIMAL(16,2)	Valor médio por venda	Numérico	0.00 a 10000.00
total_profit_generated	DECIMAL(16,2)	Lucro total gerado	Numérico	-10000.00 a 500000.00
unique_customers	BIGINT	Número de clientes únicos atendidos	Numérico	0 a 5000
top_categories_sold	ARRAY<STRING>	Categorias mais vendidas pelo funcionário	Array	["Clothing", "Accessories"]

Tabela PRODUCT_PERFORMANCE:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
product_id	STRING	Identificador único do produto	Alfanumérico	-
category	STRING	Categoria principal do produto	Categórico	"Clothing", "Accessories", "Footwear", etc.
sub_category	STRING	Subcategoria do produto	Categórico	"Shirts", "Dresses", "Pants", "Handbags", etc.
description_en	STRING	Descrição do produto em inglês	Texto	-
color	STRING	Cores disponíveis do produto	Categórico	"Black", "White", "Red", "Blue", etc.
sizes	STRING	Tamanhos disponíveis do	Categórico	"S,M,L,XL", "36,37,38,39,40",

		produto		etc.
production_cost	DECIMAL(10,2)	Custo de produção do produto	Numérico	1.00 a 500.00
times_sold	BIGINT	Número de vezes que o produto foi vendido	Numérico	0 a 10000
total_quantity_sold	BIGINT	Quantidade total vendida	Numérico	0 a 50000
avg_selling_price	DECIMAL(10,2)	Preço médio de venda	Numérico	5.00 a 1000.00
total_revenue	DECIMAL(16,2)	Receita total gerada pelo produto	Numérico	0.00 a 1000000.00
total_profit	DECIMAL(16,2)	Lucro total gerado pelo produto	Numérico	-10000.00 a 500000.00
profit_margin	DECIMAL(6,2)	Margem de lucro percentual	Numérico	-20.00 a 80.00
avg_discount_applied	DECIMAL(5,2)	Desconto médio aplicado	Numérico	0.00 a 70.00
unique_customers	BIGINT	Número de clientes únicos que compraram o produto	Numérico	0 a 5000
sold_sizes	ARRAY<STRING>	Tamanhos vendidos do produto	Array	["S", "M", "L"]

Tabela DISCOUNT_EFFECTIVENESS:

Coluna	Tipo	Descrição	Domínio	Valores Possíveis
start_date	DATE	Data de início da campanha	Data	2020-01-01 a 2023-12-31
end_date	DATE	Data de término da campanha	Data	2020-01-01 a 2023-12-31

campaign_descripiti on	STRING	Descrição da campanha de desconto	Texto	"Summer Sale", "Black Friday", etc.
category	STRING	Categoria de produto aplicável ao desconto	Categórico	"Clothing", "Accessories", "Footwear", etc.
sub_category	STRING	Subcategoria de produto aplicável ao desconto	Categórico	"Shirts", "Dresses", "Pants", "Handbags", etc.
offered_discount	DECIMAL(5,2)	Percentual de desconto oferecido	Numérico	5.00 a 70.00
total_sales	BIGINT	Número total de vendas durante a campanha	Numérico	0 a 10000
total_items_sold	BIGINT	Quantidade total de itens vendidos	Numérico	0 a 100000
revenue_during_ca mpaign	DECIMAL(16,2)	Receita total durante a campanha	Numérico	0.00 a 1000000.00
actual_avg_discoun t	DECIMAL(5,2)	Desconto médio real aplicado	Numérico	0.00 a 70.00
total_profit	DECIMAL(16,2)	Lucro total durante a campanha	Numérico	-10000.00 a 500000.00

Etapa 4 - Carga

O processo de carga dos dados foi feito utilizando o Databricks, através da própria interface do software. Uma vez carregado, os dados ficaram disponíveis para que pudessem ser acessados via múltiplas linguagens, como Python, R e SQL. Foram criadas as camadas do banco de dados e, posteriormente, as tabelas de cada camada. As tabelas da camada bronze acessam diretamente os arquivos csv que foi carregado pela interface, enquanto as camadas seguintes acessam os dados da mesma tabela da camada anterior.

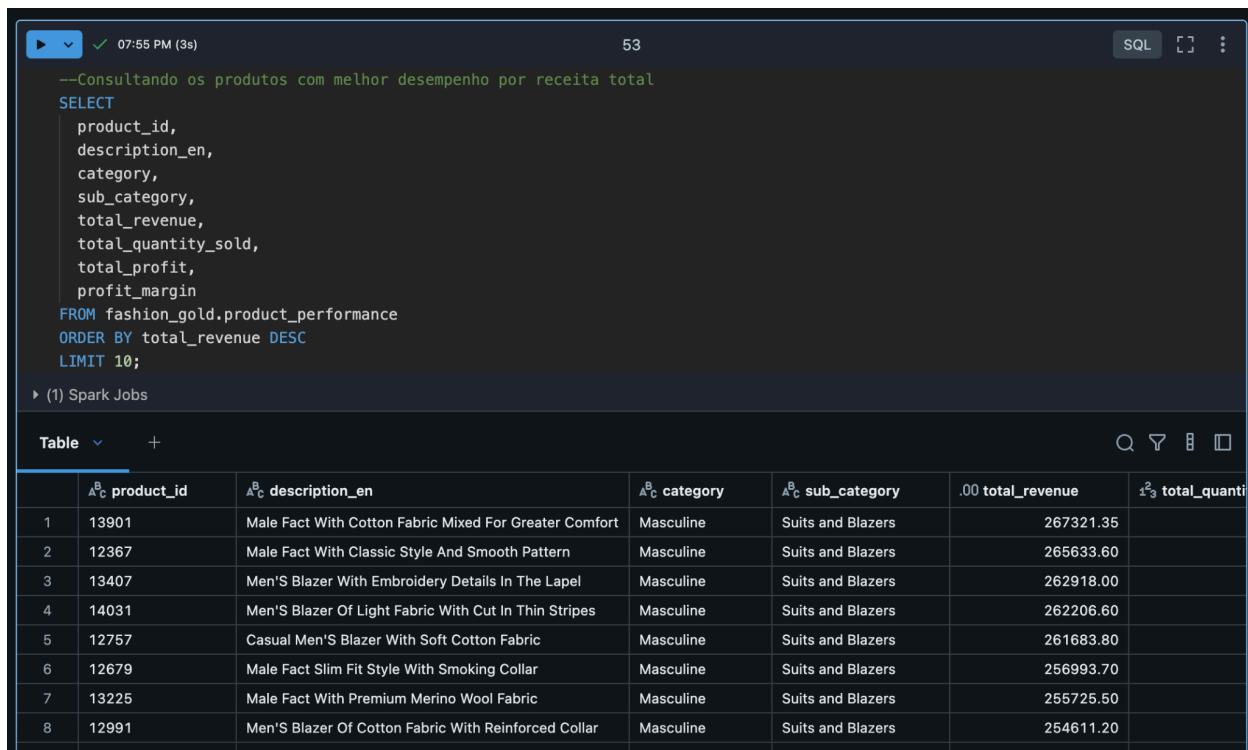
Etapa 5 - Análise

A análise qualitativa dos dados está disponível no notebook. As análises feitas não apresentaram problemas críticos nos dados, o que era previsto pela nota de usabilidade do Kaggle para o conjunto, mas ainda assim múltiplos testes foram feitos para assegurar as condições dos dados. A intenção, de forma geral, era garantir que todas as chaves primárias eram realmente chaves únicas, pois isso é necessário para garantir a integridade dos joins feitos em tabelas, garantir que não tinha um volume grande de nulos em nenhuma coluna essencial às análises (o que é comum em alguns conjuntos de dados) e garantir que todos os dados numéricos e de data estavam preenchidos com valores possíveis (não há pessoas que nasceram no futuro, produtos que custam um valor negativo, etc.). Todas as células de análise qualitativa estão com comentários sinalizando a análise feita no notebook.

Em relação às perguntas feitas no início deste projeto, temos as seguintes análises:

- Quais produtos têm melhor desempenho de vendas?

Para essa análise foram utilizadas as tabelas na camada gold que já fornecem dados tratados e enriquecidos do negócio. A tabela "product_performance" consolida informações sobre o desempenho de vendas dos produtos, unindo dados de produtos e transações da camada Silver. Ela disponibiliza métricas como volume de vendas, lucratividade, comportamento de preços e base de clientes por produto.



The screenshot shows a Jupyter Notebook cell with the following content:

```
--Consultando os produtos com melhor desempenho por receita total
SELECT
    product_id,
    description_en,
    category,
    sub_category,
    total_revenue,
    total_quantity_sold,
    total_profit,
    profit_margin
FROM fashion_gold.product_performance
ORDER BY total_revenue DESC
LIMIT 10;
```

Below the code, there is a table with the following data:

	product_id	description_en	category	sub_category	total_revenue	total_quantity_sold
1	13901	Male Fact With Cotton Fabric Mixed For Greater Comfort	Masculine	Suits and Blazers	267321.35	10
2	12367	Male Fact With Classic Style And Smooth Pattern	Masculine	Suits and Blazers	265633.60	10
3	13407	Men'S Blazer With Embroidery Details In The Lapel	Masculine	Suits and Blazers	262918.00	10
4	14031	Men'S Blazer Of Light Fabric With Cut In Thin Stripes	Masculine	Suits and Blazers	262206.60	10
5	12757	Casual Men'S Blazer With Soft Cotton Fabric	Masculine	Suits and Blazers	261683.80	10
6	12679	Male Fact Slim Fit Style With Smoking Collar	Masculine	Suits and Blazers	256993.70	10
7	13225	Male Fact With Premium Merino Wool Fabric	Masculine	Suits and Blazers	255725.50	10
8	12991	Men'S Blazer Of Cotton Fabric With Reinforced Collar	Masculine	Suits and Blazers	254611.20	10

07:55 PM (2s) 54

```
--Consultando os produtos com melhor desempenho por quantidade vendida
SELECT
    product_id,
    description_en,
    category,
    sub_category,
    total_quantity_sold,
    total_revenue,
    total_profit,
    profit_margin
FROM fashion_gold.product_performance
ORDER BY total_quantity_sold DESC
LIMIT 10;
```

▶ (1) Spark Jobs

Table +

	A _c product_id	A _c description_en	A _c category	A _c sub_category	A _c total_quantity_sold	.00 total_revenue
1	14729	Classic Turquoise With Bow	Masculine	Sweaters and Sweatshirts	1090	12974
2	12235	Short -Sleeved Men'S Shirt With Pockets	Masculine	Shirts	1076	4887
3	12417	Men'S Polo Shirt With Stripe Details	Masculine	Shirts	1070	7162
4	14458	White Jacquard Retro With Fringes	Feminine	Sweaters and Knitwear	1067	10045
5	14134	Men'S Medium Waist Jeans With Straight B...	Masculine	Pants and Jeans	1067	15724
6	13328	Light Blue Men'S Jeans With Worn Bar	Masculine	Pants and Jeans	1063	13468
7	12859	Raglan Men'S Shirt	Masculine	Shirts	1062	7493

07:55 PM (2s) 55

```
--Consultando os produtos com melhor desempenho por lucratividade
SELECT
    product_id,
    description_en,
    category,
    sub_category,
    total_profit,
    profit_margin,
    total_revenue,
    total_quantity_sold
FROM fashion_gold.product_performance
ORDER BY total_profit DESC
LIMIT 10;
```

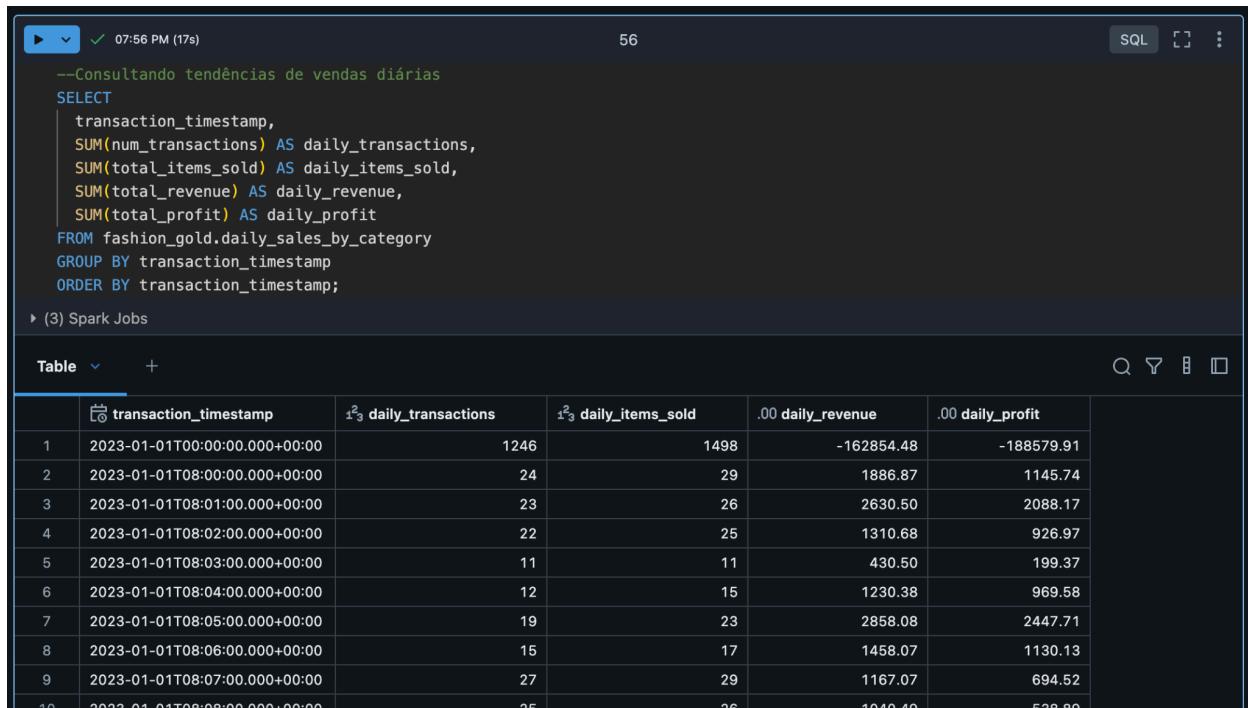
▶ (1) Spark Jobs

Table +

	A _c product_id	A _c description_en	A _c category	A _c sub_category	.00 total_profit	.00 profit_mar
1	12679	Male Fact Slim Fit Style With Smoking Collar	Masculine	Suits and Blazers	241086.90	93
2	12159	Male Fact With Detail Of Leather Appliques	Masculine	Suits and Blazers	236310.20	95
3	12601	Men'S Blazer With Thin Stripes For Fashion Look	Masculine	Suits and Blazers	235817.59	95
4	14915	Masculine Fact Of Smooth Fabric With Contemporary Design	Masculine	Suits and Blazers	231662.42	95
5	12757	Casual Men'S Blazer With Soft Cotton Fabric	Masculine	Suits and Blazers	231660.10	88
6	14031	Men'S Blazer Of Light Fabric With Cut In Thin Stripes	Masculine	Suits and Blazers	230442.24	87
7	13225	Male Fact With Premium Merino Wool Fabric	Masculine	Suits and Blazers	230253.58	90
8	12575	Masculine Twill For Elegant Casual Look	Masculine	Suits and Blazers	228461.00	92
9	12627	Slim Fit Male Fact With Tencel Fabric For Sustainability	Masculine	Suits and Blazers	228268.50	93

- Como as vendas variam ao longo do tempo?

Para essa análise, foram extraídas informações da tabela "daily_sales_by_category" na camada Gold, que agrupa dados de vendas diárias segmentados por categoria e subcategoria de produtos. A tabela combina dados das transações e produtos da camada Silver, calculando métricas como volume de vendas, receita total, lucratividade e níveis de desconto ao longo do tempo. Sua estrutura temporal permite identificar sazonalidades e tendências de vendas, importantes para o planejamento estratégico de estoque e campanhas promocionais.



```
07:56 PM (17s) 56 SQL ::

--Consultando tendências de vendas diárias
SELECT
    transaction_timestamp,
    SUM(num_transactions) AS daily_transactions,
    SUM(total_items_sold) AS daily_items_sold,
    SUM(total_revenue) AS daily_revenue,
    SUM(total_profit) AS daily_profit
FROM fashion_gold.daily_sales_by_category
GROUP BY transaction_timestamp
ORDER BY transaction_timestamp;
```

(3) Spark Jobs

Table +

	transaction_timestamp	daily_transactions	daily_items_sold	daily_revenue	daily_profit
1	2023-01-01T00:00:00.000+00:00	1246	1498	-162854.48	-188579.91
2	2023-01-01T08:00:00.000+00:00	24	29	1886.87	1145.74
3	2023-01-01T08:01:00.000+00:00	23	26	2630.50	2088.17
4	2023-01-01T08:02:00.000+00:00	22	25	1310.68	926.97
5	2023-01-01T08:03:00.000+00:00	11	11	430.50	199.37
6	2023-01-01T08:04:00.000+00:00	12	15	1230.38	969.58
7	2023-01-01T08:05:00.000+00:00	19	23	2858.08	2447.71
8	2023-01-01T08:06:00.000+00:00	15	17	1458.07	1130.13
9	2023-01-01T08:07:00.000+00:00	27	29	1167.07	694.52
10	2023-01-01T08:08:00.000+00:00	25	26	1040.49	522.80

07:56 PM (10s) 57 SQL ⌂ ⌂

```
--Consultando tendências de vendas por mês
SELECT
    YEAR(transaction_timestamp) AS year,
    MONTH(transaction_timestamp) AS month,
    SUM(num_transactions) AS monthly_transactions,
    SUM(total_items_sold) AS monthly_items_sold,
    SUM(total_revenue) AS monthly_revenue,
    SUM(total_profit) AS monthly_profit
FROM fashion_gold.daily_sales_by_category
GROUP BY YEAR(transaction_timestamp), MONTH(transaction_timestamp)
ORDER BY year, month;
```

(2) Spark Jobs

Table +

	year	month	monthly_transactions	monthly_items_sold	.00 monthly_revenue	.00 monthly_profit
1	2023	1	166604	188884	13901921.78	10550503.86
2	2023	2	116389	131323	11344209.79	9059417.61
3	2023	3	255162	290644	30245575.17	25279487.35
4	2023	4	182388	206240	23409291.97	19743961.77
5	2023	5	206926	234227	24079336.01	20101832.51
6	2023	6	128943	145290	18467461.66	15877706.43
7	2023	7	143948	162390	17215818.80	14330383.11
8	2023	8	125120	141130	16166374.30	13654204.31

07:58 PM (1s) 59 SQL ⌂ ⌂

```
--Consultando campanhas por ROI (retorno sobre investimento em desconto)
SELECT
    campaign_description,
    category,
    sub_category,
    offered_discount,
    actual_avg_discount,
    total_sales,
    revenue_during_campaign,
    total_profit,
    total_profit / (revenue_during_campaign * actual_avg_discount / 100) AS estimated_roi
FROM fashion_gold.discount_effectiveness
WHERE actual_avg_discount > 0
ORDER BY estimated_roi DESC;
```

(1) Spark Jobs

Table +

	campaign_description	category	sub_category	.00 offered_discount	.00 actual_avg_discou
1	20% discount during our Autumn Essentials Sale	Masculine	Sportswear	0.20	0.20
2	20% discount during our Autumn Essentials Sale	Feminine	Sportswear	0.20	0.20
3	20% discount during our Autumn Essentials Sale	Masculine	Sweaters and Sweatshirts	0.20	0.20
4	20% discount during our Autumn Essentials Sale	Masculine	Sweaters and Sweatshirts	0.20	0.20
5	20% discount during our Autumn Essentials Sale	Children	Sweaters	0.20	0.20
6	20% discount during our Autumn Essentials Sale	Masculine	Sportswear	0.20	0.20
7	20% discount during our Autumn Essentials Sale	Feminine	Sweaters and Knitwear	0.20	0.20
8	20% discount during our Autumn Essentials Sale	Children	T-shirts	0.20	0.20

- Qual é a eficácia de diferentes campanhas promocionais?

Para essa análise, foram extraídas informações da tabela "store_performance" na camada gold, que calcula métricas de desempenho por loja a partir de dados da camada Silver. A tabela integra informações de lojas, transações e produtos, fornecendo indicadores como receita total, lucro, margem de lucro e eficiência operacional por loja. Sua estrutura permite

comparações entre diferentes unidades e regiões, identificando oportunidades de melhoria.

07:57 PM (2s) 58

```
--Consultando campanhas por receita gerada
SELECT
    campaign_description,
    category,
    sub_category,
    start_date,
    end_date,
    DATEDIFF(end_date, start_date) AS campaign_duration_days,
    offered_discount,
    actual_avg_discount,
    total_sales,
    total_items_sold,
    revenue_during_campaign,
    revenue_during_campaign / DATEDIFF(end_date, start_date) AS daily_revenue,
    total_profit,
    (total_profit / revenue_during_campaign) * 100 AS profit_margin
FROM fashion_gold.discount_effectiveness
ORDER BY revenue_during_campaign DESC;
```

(2) Spark Jobs

Table +

	^A _C campaign_description	^A _C category	^A _C sub_category	^A _T start_date	^A _T end_date	^I ₂ ₃ camp
1	35% discount during our Early Spring Collection Refresh	Children	Girl and Boy (1-5 years, 6-14 years)	2024-03-15	2024-03-31	
2	45% discount during our Fall Collection Launch	Masculine	Pants and Jeans	2024-09-01	2024-09-15	
3	35% discount during our Early Spring Collection Refresh	Feminine	Dresses and Jumpsuits	2024-03-15	2024-03-31	
4	35% discount during our Early Spring Collection Refresh	Children	Girl and Boy (1-5 years, 6-14 years)	2023-03-15	2023-03-31	
5	25% discount during our Mid-Spring Refresh Sale	Masculine	Pants and Jeans	2024-05-01	2024-05-15	

07:58 PM (1s) 59

SQL ☰ ⚙

```
--Consultando campanhas por ROI (retorno sobre investimento em desconto)
SELECT
    campaign_description,
    category,
    sub_category,
    offered_discount,
    actual_avg_discount,
    total_sales,
    revenue_during_campaign,
    total_profit,
    total_profit / (revenue_during_campaign * actual_avg_discount / 100) AS estimated_roi
FROM fashion_gold.discount_effectiveness
WHERE actual_avg_discount > 0
ORDER BY estimated_roi DESC;
```

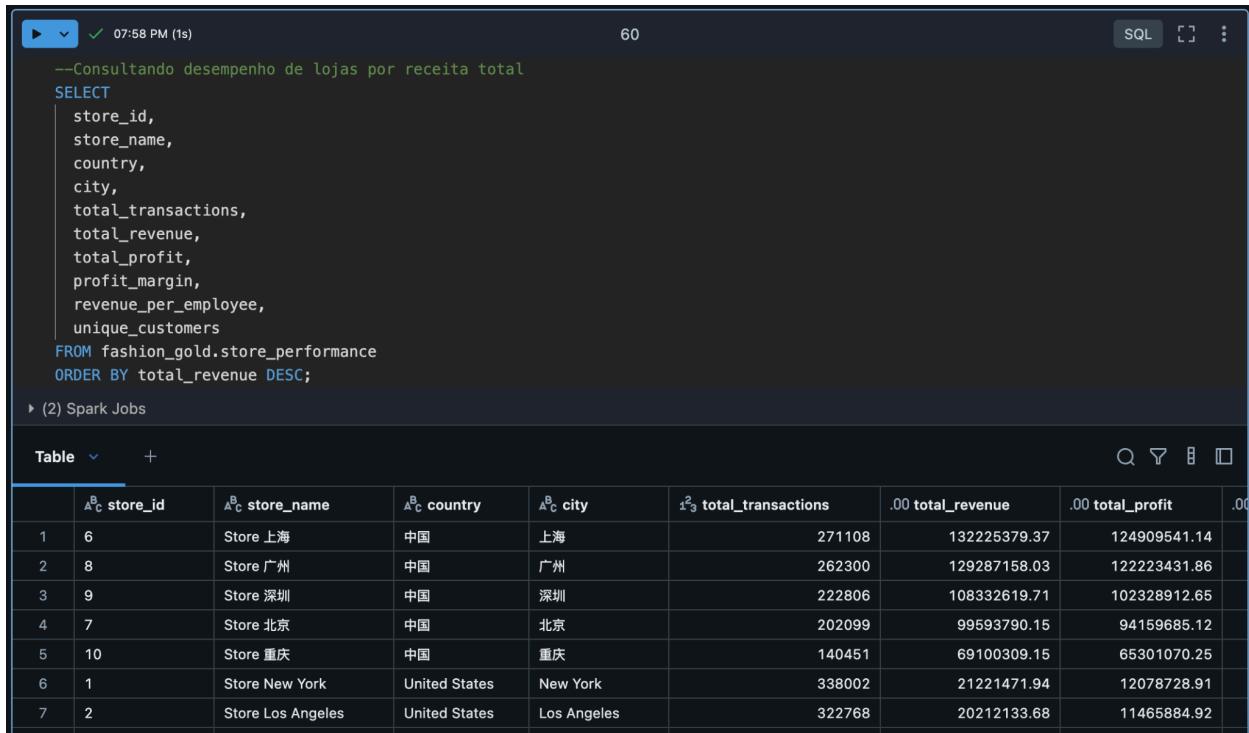
(1) Spark Jobs

Table +

	^A _C campaign_description	^A _C category	^A _C sub_category	.00 offered_discount	.00 actual_avg_discou
1	20% discount during our Autumn Essentials Sale	Masculine	Sportswear	0.20	0.2
2	20% discount during our Autumn Essentials Sale	Feminine	Sportswear	0.20	0.2
3	20% discount during our Autumn Essentials Sale	Masculine	Sweaters and Sweatshirts	0.20	0.2
4	20% discount during our Autumn Essentials Sale	Masculine	Sweaters and Sweatshirts	0.20	0.2
5	20% discount during our Autumn Essentials Sale	Children	Sweaters	0.20	0.2
6	20% discount during our Autumn Essentials Sale	Masculine	Sportswear	0.20	0.2
7	20% discount during our Autumn Essentials Sale	Feminine	Sweaters and Knitwear	0.20	0.2
8	20% discount during our Autumn Essentials Sale	Children	Accessories	0.20	0.2
9	20% discount during our Autumn Essentials Sale	Children	Accessories	0.20	0.2

- Qual é o desempenho comparativo entre diferentes lojas?

Para essa análise, foram utilizados dados da tabela "store_performance" na camada gold, que apresenta métricas do desempenho de cada loja. Esta tabela combina informações das tabelas de lojas, transações e produtos da camada Silver, calculando indicadores como volume de vendas, receita total, lucratividade e eficiência por funcionário. A estrutura da tabela permite análises comparativas entre lojas e regiões, importantes para identificar unidades de alto desempenho e orientar decisões estratégicas de expansão ou otimização operacional da rede.



The screenshot shows a SQL query being run in a terminal or IDE. The query retrieves data from the 'store_performance' table in the 'fashion_gold' database, specifically ordering by total_revenue in descending order. The results are displayed in a table format below the query.

```
--Consultando desempenho de lojas por receita total
SELECT
    store_id,
    store_name,
    country,
    city,
    total_transactions,
    total_revenue,
    total_profit,
    profit_margin,
    revenue_per_employee,
    unique_customers
FROM fashion_gold.store_performance
ORDER BY total_revenue DESC;
```

	store_id	store_name	country	city	total_transactions	total_revenue	total_profit
1	6	Store 上海	中国	上海	271108	132225379.37	124909541.14
2	8	Store 广州	中国	广州	262300	129287158.03	122223431.86
3	9	Store 深圳	中国	深圳	222806	108332619.71	102328912.65
4	7	Store 北京	中国	北京	202099	99593790.15	94159685.12
5	10	Store 重庆	中国	重庆	140451	69100309.15	65301070.25
6	1	Store New York	United States	New York	338002	21221471.94	12078728.91
7	2	Store Los Angeles	United States	Los Angeles	322768	20212133.68	11465884.92

▶ ✓ 07:59 PM (1s) 61

```
--Consultando lojas por eficiência (receita por funcionário)
SELECT
    store_id,
    store_name,
    country,
    city,
    number_of_employees,
    total_revenue,
    revenue_per_employee,
    total_profit,
    profit_margin
FROM fashion_gold.store_performance
ORDER BY revenue_per_employee DESC;
```

▶ (1) Spark Jobs

Table +

	store_id	store_name	country	city	number_of_employees	total_revenue	revenue_per_employee
1	6	Store 上海	中国	上海	8	132225379.37	16528172.4212
2	8	Store 广州	中国	广州	10	129287158.03	12928715.8030
3	9	Store 深圳	中国	深圳	9	108332619.71	12036957.7455
4	7	Store 北京	中国	北京	10	99593790.15	99593790.1500
5	10	Store 重庆	中国	重庆	10	69100309.15	6910030.9150
6	2	Store Los Angeles	United States	Los Angeles	8	20212133.68	2526516.7100
7	1	Store New York	United States	New York	10	21221471.94	2122147.1940
8	4	Store Houston	United States	Houston	10	13918817.94	1391881.7940
9	11	Store Berlin	Deutschland	Berlin	9	10881034.07	1209003.7855

▶ ✓ 07:59 PM (1s) 62

```
--Consultando desempenho de lojas por receita total
SELECT
    store_id,
    store_name,
    country,
    city,
    total_transactions,
    total_revenue,
    total_profit,
    profit_margin,
    revenue_per_employee,
    unique_customers
FROM fashion_gold.store_performance
ORDER BY total_revenue DESC;
```

▶ (1) Spark Jobs

Table +

	store_id	store_name	country	city	total_transactions	total_revenue	total_profit	unique_customers
1	6	Store 上海	中国	上海	271108	132225379.37	124909541.14	132225379.37
2	8	Store 广州	中国	广州	262300	129287158.03	122223431.86	129287158.03
3	9	Store 深圳	中国	深圳	222806	108332619.71	102328912.65	108332619.71
4	7	Store 北京	中国	北京	202099	99593790.15	94159685.12	99593790.15
5	10	Store 重庆	中国	重庆	140451	69100309.15	65301070.25	69100309.15
6	1	Store New York	United States	New York	338002	21221471.94	12078728.91	21221471.94
7	2	Store Los Angeles	United States	Los Angeles	322768	20212133.68	11465884.92	20212133.68
8	4	Store Houston	United States	Houston	220249	13918817.94	7969689.37	13918817.94

- Qual é o desempenho comparativo entre diferentes regiões?

Para essa análise, foram coletados dados da tabela "customer_profile" na camada gold, que mede o perfil e comportamento de compra de cada cliente. Esta tabela une informações das tabelas de clientes, transações e produtos da camada silver, gerando métricas como frequência de compra, valor médio gasto, categorias preferidas e recência de compras. A estrutura permite segmentação e análise demográfica dos clientes, importantes para personalização de campanhas de marketing e desenvolvimento de estratégias de fidelização baseadas no comportamento real de consumo.

▶ ✓ 07:59 PM (1s) 63

```
--Consultando desempenho por país
SELECT
    country,
    COUNT(*) AS num_stores,
    SUM(total_transactions) AS total_transactions,
    SUM(total_revenue) AS total_revenue,
    SUM(total_profit) AS total_profit,
    SUM(total_profit) / SUM(total_revenue) * 100 AS overall_profit_margin,
    SUM(unique_customers) AS total_unique_customers,
    SUM(total_revenue) / COUNT(*) AS avg_revenue_per_store
FROM fashion_gold.store_performance
GROUP BY country
ORDER BY total_revenue DESC;
```

▶ (2) Spark Jobs

Table +

	country	num_stores	total_transactions	total_revenue	total_profit	overall_profit_margin	total_unique_customers
1	中国	5	1098764	538539256.41	508922641.02	94.500600	1098764
2	United States	5	1194228	75023511.70	42741231.41	56.970400	1194228
3	Deutschland	5	530336	29842670.41	15528031.90	52.033000	530336
4	France	5	458571	25838685.73	13507230.64	52.275200	458571
5	Portugal	5	410739	23223036.14	12188387.06	52.484000	410739
6	Espanha	5	404481	22878360.15	12026212.73	52.565900	404481
7	United Kingdom	5	443285	17399354.45	5458318.66	31.370800	443285

▶ ✓ 08:00 PM (1s) 64

```
--Top10 cidades mais lucrativas
SELECT
    country,
    city,
    COUNT(*) AS num_stores,
    SUM(total_transactions) AS total_transactions,
    SUM(total_revenue) AS total_revenue,
    SUM(total_profit) AS total_profit,
    SUM(total_profit) / SUM(total_revenue) * 100 AS overall_profit_margin
FROM fashion_gold.store_performance
GROUP BY country, city
ORDER BY total_revenue DESC
LIMIT 10;
```

▶ (2) Spark Jobs

Table +

	country	city	num_stores	total_transactions	total_revenue	total_profit	overall_profit_margin
1	中国	上海	1	271108	132225379.37	124909541.14	94.467100
2	中国	广州	1	262300	129287158.03	122223431.86	94.536400
3	中国	深圳	1	222806	108332619.71	102328912.65	94.458100
4	中国	北京	1	202099	99593790.15	94159685.12	94.543700
5	中国	重庆	1	140451	69100309.15	65301070.25	94.501800
6	United States	New York	1	338002	21221471.94	12078728.91	56.917500
7	United States	Los Angeles	1	322768	20212133.68	11465884.92	56.727700
8	United States	Houston	1	220249	13918817.94	7969689.37	57.258400
9	Deutschland	Berlin	1	192712	10881034.07	5690793.56	52.300100

- Quais produtos e categorias geram maiores margens de lucro?

Para essa análise, foram consultados novamente os dados da tabela "product_performance" na camada Gold, que oferece uma visão abrangente do desempenho comercial de cada item. A análise foi feita para consultar as categorias e produtos que tem maior margem de lucro (o que é calculado comparando o custo à receita).

▶ ✅ 08:00 PM (3s) 65 SQL ⚙️ ⚖️

```
--Consultando categorias por margem de lucro
SELECT
    category,
    sub_category,
    COUNT(*) AS num_products,
    SUM(total_quantity_sold) AS total_quantity_sold,
    SUM(total_revenue) AS total_revenue,
    SUM(total_profit) AS total_profit,
    SUM(total_profit) / SUM(total_revenue) * 100 AS category_profit_margin
FROM fashion_gold.product_performance
GROUP BY category, sub_category
ORDER BY category_profit_margin DESC;
```

▶ (2) Spark Jobs

Table +

	category	sub_category	num_products	total_quantity_sold	total_revenue	total_profit
1	Masculine	Sportswear	690	338381	36655000.20	31231695.33
2	Masculine	Suits and Blazers	690	319742	75986464.65	64583852.05
3	Masculine	T-shirts and Polos	690	371025	30679684.20	26053641.96
4	Masculine	Coats and Blazers	690	336147	54563778.08	46333051.47
5	Children	Baby (0-12 months)	690	104939	7050798.10	5979294.47
6	Masculine	Sweaters and Sweatshirts	690	356515	36619413.24	30999699.78
7	Masculine	Pants and Jeans	690	371768	51652271.41	43432796.10
8	Masculine	Underwear and Pajamas	690	158592	10537322.10	8857590.75
9	Children	Accessories	690	137501	5348536.15	444444.44

▶ ✅ 08:00 PM (3s) 66 Microsoft Excel

```
--Top20 produtos mais rentáveis (maior margem de lucro)
SELECT
    product_id,
    description_en,
    category,
    sub_category,
    production_cost,
    avg_selling_price,
    total_quantity_sold,
    total_revenue,
    total_profit,
    profit_margin
FROM fashion_gold.product_performance
WHERE total_quantity_sold > 10
ORDER BY profit_margin DESC
LIMIT 20;
```

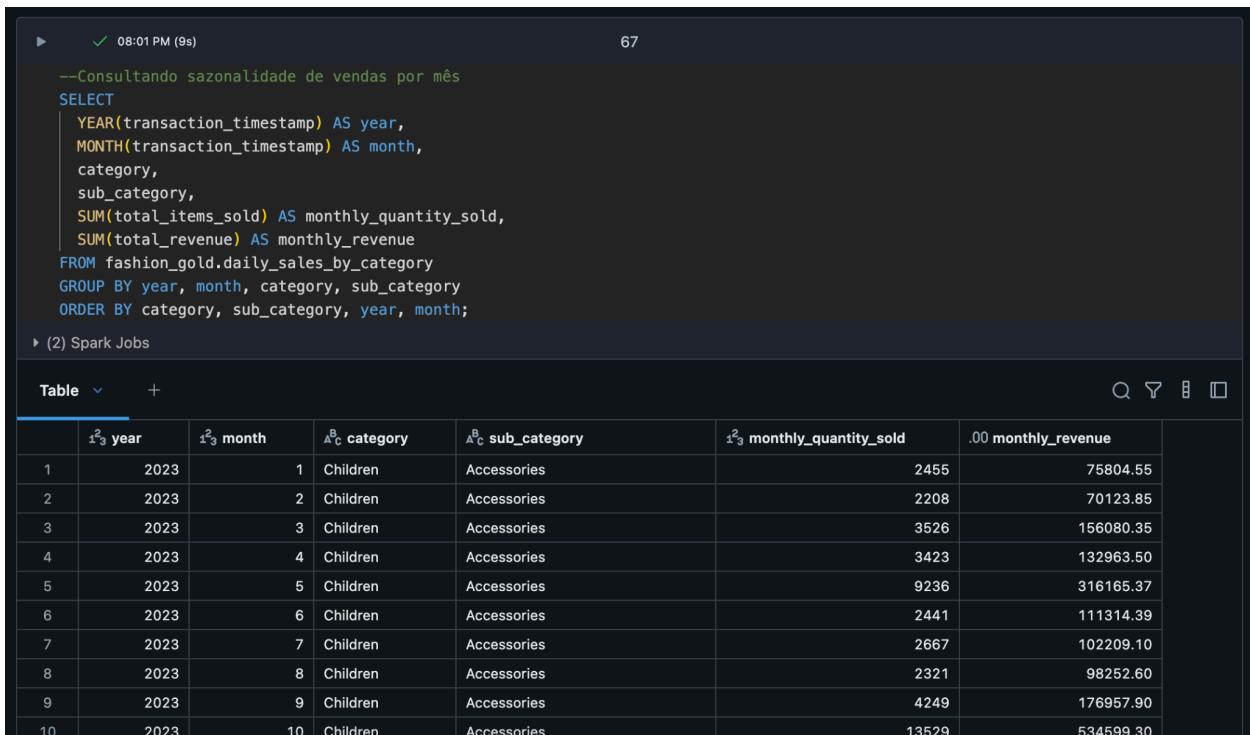
▶ (1) Spark Jobs

Table +

	product_id	description_en	category	sub_category	production_cost	avg_selling_price
1	17185	Children'S Pajama With Long Sleeve And Glitter Details	Children	Pajamas	3.47	3.47
2	17696	Men'S Colorful Pants	Masculine	Pants and Jeans	4.31	4.31
3	16416	Cotton Pajamas With Animal Print	Feminine	Lingerie and Pajamas	2.06	2.06
4	17857	Casual Seda Pink With Bow	Children	Baby (0-12 months)	2.82	2.82
5	15049	Sports Borde Seda Padded	Children	Baby (0-12 months)	5.37	5.37

- Podemos manter um estoque dinâmico de produtos com base na sazonalidade do negócio?

A análise visa entender quais categorias e subcategorias de produtos são os mais vendidos em cada mês de cada ano, tanto em quantidade quanto em receita. A análise é útil para entendermos a necessidade de aumentar o estoque de um produto específico em um período específico do ano, e também é necessária para entendermos se o comportamento foi exclusivo de apenas um dos anos mas não se repetiu nos demais.



```
--Consultando sazonalidade de vendas por mês
SELECT
    YEAR(transaction_timestamp) AS year,
    MONTH(transaction_timestamp) AS month,
    category,
    sub_category,
    SUM(total_items_sold) AS monthly_quantity_sold,
    SUM(total_revenue) AS monthly_revenue
FROM fashion_gold.daily_sales_by_category
GROUP BY year, month, category, sub_category
ORDER BY category, sub_category, year, month;
```

	year	month	category	sub_category	monthly_quantity_sold	monthly_revenue
1	2023	1	Children	Accessories	2455	75804.55
2	2023	2	Children	Accessories	2208	70123.85
3	2023	3	Children	Accessories	3526	156080.35
4	2023	4	Children	Accessories	3423	132963.50
5	2023	5	Children	Accessories	9236	316165.37
6	2023	6	Children	Accessories	2441	111314.39
7	2023	7	Children	Accessories	2667	102209.10
8	2023	8	Children	Accessories	2321	98252.60
9	2023	9	Children	Accessories	4249	176957.90
10	2023	10	Children	Accessories	13529	534599.30

```

--Pico de meses por categoria de produto
WITH monthly_sales AS (
    SELECT
        MONTH(transaction_timestamp) AS month,
        category,
        SUM(total_items_sold) AS monthly_quantity_sold
    FROM fashion_gold.daily_sales_by_category
    GROUP BY month, category
),
category_totals AS (
    SELECT
        category,
        SUM(monthly_quantity_sold) AS total_quantity_sold
    FROM monthly_sales
    GROUP BY category
)
SELECT
    ms.month,
    ms.category,
    ms.monthly_quantity_sold,
    ms.monthly_quantity_sold / ct.total_quantity_sold * 100 AS percentage_of_annual_sales,
    RANK() OVER (PARTITION BY ms.category ORDER BY ms.monthly_quantity_sold DESC) AS month_rank
FROM monthly_sales ms
JOIN category_totals ct ON ms.category = ct.category
ORDER BY ms.category, month_rank;

```

(5) Spark Jobs

Table +

month	category	monthly_quantity_sold	percentage_of_annual_sales	month_rank
1	Electronics	1000	100.0	1
2	Electronics	800	80.0	2
3	Electronics	600	60.0	3
4	Electronics	400	40.0	4
5	Electronics	200	20.0	5
6	Electronics	100	10.0	6
7	Electronics	50	5.0	7
8	Electronics	30	3.0	8
9	Electronics	20	2.0	9
10	Electronics	10	1.0	10
11	Electronics	5	0.5	11
12	Electronics	3	0.3	12
1	Apparel	800	80.0	13
2	Apparel	600	60.0	14
3	Apparel	400	40.0	15
4	Apparel	200	20.0	16
5	Apparel	100	10.0	17
6	Apparel	50	5.0	18
7	Apparel	30	3.0	19
8	Apparel	20	2.0	20
9	Apparel	10	1.0	21
10	Apparel	5	0.5	22
11	Apparel	3	0.3	23
12	Apparel	2	0.2	24
1	Home & Garden	600	60.0	25
2	Home & Garden	400	40.0	26
3	Home & Garden	200	20.0	27
4	Home & Garden	100	10.0	28
5	Home & Garden	50	5.0	29
6	Home & Garden	30	3.0	30
7	Home & Garden	20	2.0	31
8	Home & Garden	10	1.0	32
9	Home & Garden	5	0.5	33
10	Home & Garden	3	0.3	34
11	Home & Garden	2	0.2	35
12	Home & Garden	1	0.1	36
1	Food & Beverage	400	40.0	37
2	Food & Beverage	200	20.0	38
3	Food & Beverage	100	10.0	39
4	Food & Beverage	50	5.0	40
5	Food & Beverage	30	3.0	41
6	Food & Beverage	20	2.0	42
7	Food & Beverage	10	1.0	43
8	Food & Beverage	5	0.5	44
9	Food & Beverage	3	0.3	45
10	Food & Beverage	2	0.2	46
11	Food & Beverage	1	0.1	47
12	Food & Beverage	0.5	0.5	48

- Quais são os padrões de comportamento dos clientes?

A análise usa da tabela customer_profile para definir 4 perfis de clientes: VIP, Regular, Recente e Ocasional. Cada cliente tem um padrão de comportamento e gasta um valor por compra, portanto a ideia é que esse padrão seja captado, e isso pode ajudar em campanhas de marketing e ações de retenção e aumento do ticket médio.

```

Just now (6s) 69
--Clientes por frequência e valor
SELECT
CASE
    WHEN total_purchases >= 5 AND total_spent >= 1000 THEN 'VIP'
    WHEN total_purchases >= 3 AND total_spent >= 500 THEN 'Regular'
    WHEN days_since_last_purchase <= 90 THEN 'Recente'
    ELSE 'Ocasional'
END AS customer_segment,
COUNT(*) AS segment_size,
AVG(total_purchases) AS avg_purchases,
AVG(total_spent) AS avg_spend,
AVG(avg_purchase_value) AS avg_order_value,
SUM(total_spent) AS segment_total_spend,
SUM(total_spent) / SUM(total_purchases) AS segment_avg_order_value
FROM fashion_gold.customer_profile
GROUP BY customer_segment
ORDER BY avg_spend DESC;

```

(2) Spark Jobs

	customer_segment	segment_size	avg_purchases	avg_spend	avg_order_value	segment_total_spend
1	VIP	88577	7.9318671889994015	3517.282662	331.2955871023	311550346.33
2	Regular	149560	5.73035571008291	1243.693696	221.7172755141	186006829.17
3	Recente	155473	4.056620763733896	243.558008	58.6926281748	37866694.14
4	Ocasional	1249696	1.880534145904284	221.684834	83.0669579948	197321005.35

- Qual o perfil dos nossos clientes em relação a idade, sexo?

Também temos aqui uma análise que, além de nos fazer entender o perfil do cliente, permite a melhoria e direcionamento adequado de campanhas de marketing, por exemplo. As análises foram feitas gênero e faixa etária (devido a quantidade de idades, a definição de faixas torna a análise mais simples).

Just now (7s) 70 SQL ⚙️ ⚖️

```
--Análise de clientes por gênero
SELECT
    gender,
    COUNT(*) AS customer_count,
    COUNT(*) * 100.0 / (SELECT COUNT(*) FROM fashion_gold.customer_profile) AS percentage,
    AVG(age) AS avg_age,
    AVG(total_purchases) AS avg_purchases,
    AVG(total_spent) AS avg_total_spent,
    SUM(total_spent) AS total_revenue
FROM fashion_gold.customer_profile
GROUP BY gender
ORDER BY customer_count DESC;
```

▶ (6) Spark Jobs

Table +

	gender	customer_count	percentage	avg_age	avg_purchases	avg_total_spent	total_revenue
1	M	964562	58.69643267900196	31.221716177912878	1.6957458411175228	428.574877	29417
2	F	677041	41.19993476564925	31.886554876292575	4.2861938937228325	734.959410	43810
3	D	1703	0.10363255534879	31.424544920728128	1.6576629477392837	382.986086	46

↓ 3 rows | 6.79s runtime Refreshed now

08:03 PM (5s) 71

```
--Análise de clientes por faixa etária
SELECT
    CASE
        WHEN age < 18 THEN 'Under 18'
        WHEN age BETWEEN 18 AND 24 THEN '18-24'
        WHEN age BETWEEN 25 AND 34 THEN '25-34'
        WHEN age BETWEEN 35 AND 44 THEN '35-44'
        WHEN age BETWEEN 45 AND 54 THEN '45-54'
        WHEN age BETWEEN 55 AND 64 THEN '55-64'
        WHEN age >= 65 THEN '65+'
        ELSE 'Unknown'
    END AS age_group,
    COUNT(*) AS customer_count,
    COUNT(*) * 100.0 / (SELECT COUNT(*) FROM fashion_gold.customer_profile) AS percentage,
    AVG(total_purchases) AS avg_purchases,
    AVG(total_spent) AS avg_total_spent
FROM fashion_gold.customer_profile
GROUP BY age_group;
```

▶ (6) Spark Jobs

Table +

	age_group	customer_count	percentage	avg_purchases	avg_total_spent
1	45-54	162923	9.91434340287202	2.7770664669813345	547.244229
2	55-64	72477	4.41043846976765	2.669426162782676	727.423800
3	35-44	339163	20.63906539621957	2.8674236281669874	390.821980
4	25-34	409723	24.92379702872417	2.7523002722708087	416.086477

Autoavaliação

Em relação aos objetivos iniciais do trabalho, fiquei muito feliz e satisfeita por encontrar um conjunto de dados que se encaixou quase que perfeitamente aos objetivos, dado que esse provavelmente era um grande obstáculo inicial e poderia mudar todo o rumo do projeto. Embora já trabalhe com dados há alguns anos, não tenho tanta experiência em análise de dados de plataforma de vendas - trabalho, hoje, com produtização de dados financeiros -, e esse conhecimento me parece ser algo útil para o mercado de trabalho. Também nunca tinha tido experiência com o databricks, que também é uma ferramenta bastante popular no mercado.

No geral as perguntas puderam ser respondidas com as análises feitas, e os dados foram carregados e modelados de forma eficiente na plataforma. Em minha própria opinião sobre o projeto, a qualidade da parte técnica está alta e atendeu às minhas expectativas - consegui criar as camadas, verificar uma variedade grande de possíveis problemas existentes no conjunto de dados e criar métricas pertinentes e necessárias para responder às perguntas. Quanto ao relatório, acredito que iniciei bem mas houve uma queda na qualidade das explicações por conta de alguns problemas grandes da empresa que aconteceram enquanto eu era a única disponível no time, e acabaram me tomado muito tempo nas últimas duas semanas.

Acredito que o objetivo de entender melhor a parte de engenharia de dados - visto que trabalho com análise e machine learning e entendia somente o necessário da parte de infraestrutura - também foi cumprido com sucesso.