



Indicium

Lívia Nobre de Mesquita

RELATÓRIO DO DESAFIO LIGHTHOUSE -
PIPELINE DA ADVENTURE WORKS



Fortaleza

Janeiro 2025

Sumário

1.	Introdução	4
2.	Objetivo	5
3.	Modelagem Conceitual e EDA	6
4.	Data Warehouse	10
4.1.	Big Query	10
5.	dbt Cloud	12
5.1.	Conexão com Big Query	12
5.2.	Conexão dbt e Github	13
5.3.	Transformações dos dados	14
5.3.1.	dbt_project.yml	14
5.3.2.	Staging	14
5.3.3.	Marts	16
5.4.	Ambiente de Produção	17
6.	Orquestração com Apache Airflow	18
1.	Instalação do CLI Astro	18
2.	Instalação do Docker	18
3.	Clonagem do Repositório DBT Cloud	18
4.	Conexão do DBT Cloud com o Airflow	19
5.	Criação e Execução de um Job no DBT Cloud	20
6.	Criação da DAG no Airflow	21
7.	Verificação da Execução da DAG	21
8.	Apresentação dos dados	23
8.1.	Mockup	23
8.2.	Conexão do Power BI com a BQ	25
8.3.	Relatório Geral para o CEO	26
8.3.1.	Métricas	27
8.3.2.	Gráficos	27
8.4.	Dashboard	29
8.4.1.	Visão Geral	29
8.4.2.	Pedido	33

8.4.3.	Região	36
8.4.4.	Cliente	39
9.	Advanced Analytics.....	42
9.1.	Carregamento dos Dados.....	42
9.2.	Carregamento, processamento e transformação dos dados.....	43
9.3.	Modelagem e Redefinição de Índices.....	43
9.4.	Questões e Respostas	43
9.4.1.	Questão 8.....	43
9.4.2.	Questão 9.....	44
9.4.3.	Questão 10.....	44
9.4.4.	Questão 11	45
10.	Conclusão	46

1. Introdução

O crescimento acelerado da Adventure Works trouxe a necessidade de uma visão mais estratégica e dinâmica sobre seus setores e desempenho, de forma a facilitar a tomada de decisões. Nesse cenário, os dados se tornam essenciais para compreender os processos e identificar oportunidades de melhoria. Este projeto visa implementar uma abordagem moderna de analytics, alinhada às melhores práticas do mercado, para potencializar decisões estratégicas e destacar a empresa.

Neste documento, serão apresentados os passos para construir um pipeline completo de dados, seguindo os princípios do Modern Data Analytics. O projeto abrange desde a criação de um data warehouse (DW) até análises avançadas e visualizações que geram insights valiosos, tanto para a operação quanto para a gestão estratégica da Adventure Works.

2. Objetivo

O objetivo deste projeto é transformar os dados existentes da Adventure Works em insights estratégicos, utilizando a implementação de um data warehouse (DW) e ferramentas modernas de Analytics. A iniciativa busca não apenas organizar e centralizar as informações, mas também oferecer uma base sólida para análises que orientem decisões mais assertivas e direcionem o crescimento da empresa.

Para a concretização do projeto, será criada uma infraestrutura de dados robusta e escalável, capaz de atender às necessidades atuais e futuras da organização. Isso inclui o desenvolvimento de tabelas de fato e dimensões que respondam a perguntas de negócio essenciais, a construção de visualizações claras e intuitivas para suporte a decisões tanto operacionais quanto gerenciais e, por fim, a aplicação de análises preditivas para um planejamento mais eficiente da demanda. Além disso, será considerado o alinhamento com as melhores práticas de governança de dados, garantindo segurança, integridade e acessibilidade da informação em toda a empresa.

3. Modelagem Conceitual e EDA

A Adventure Works possui um banco de dados transacional que armazena informações provenientes de diferentes áreas da empresa. Esse banco de dados é composto por 68 tabelas distribuídas em cinco schemas principais: HR (Recursos Humanos), Sales (Vendas), Production (Produção) e Purchasing (Compras).

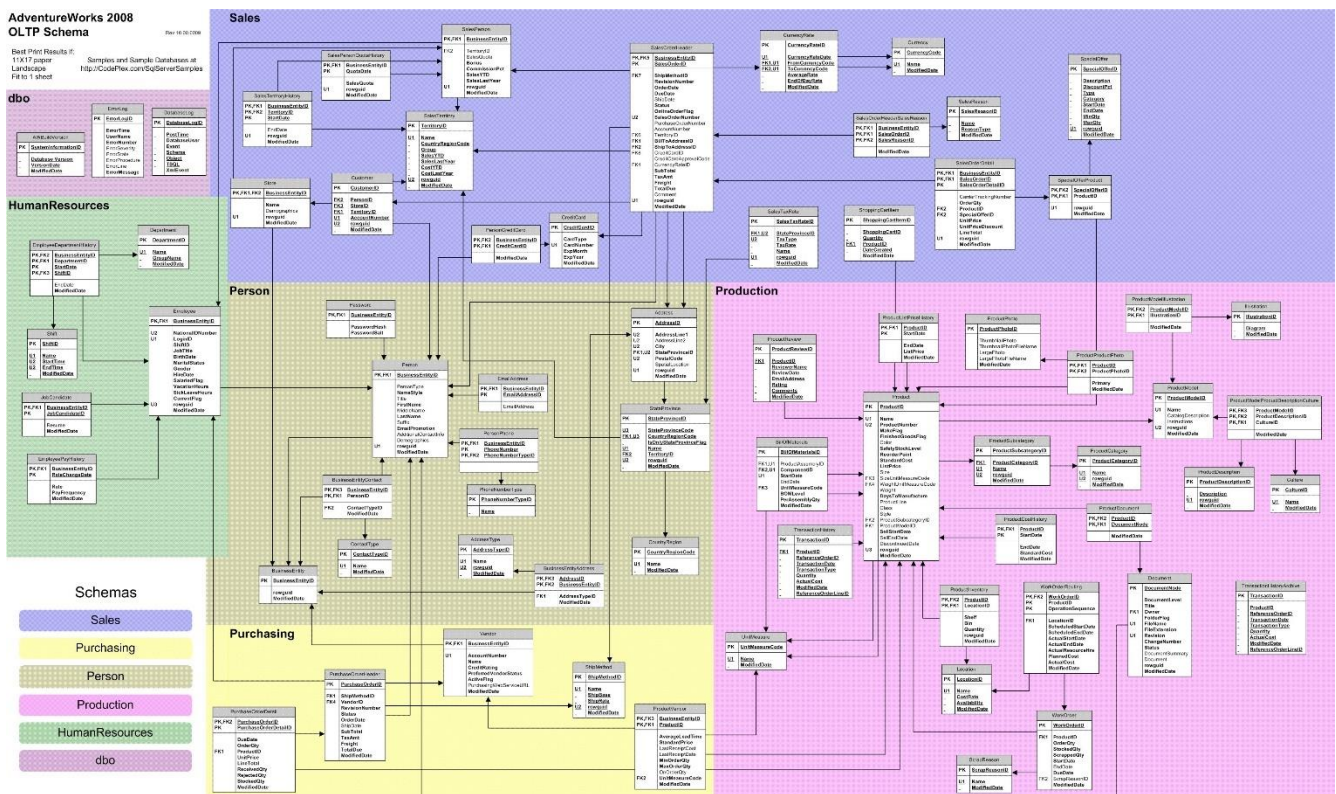


Diagrama completo dos dados da Adventure Works.

Para iniciar o projeto de forma eficiente, é essencial realizar um mapeamento detalhado dos dados disponíveis e identificar as principais perguntas de negócio que precisam ser respondidas. Essa abordagem garante que a modelagem dos dados esteja alinhada aos objetivos estratégicos, utilizando apenas as informações necessárias para gerar insights. As perguntas de negócio foram estruturadas em torno de três recortes principais relacionados às vendas:

- Por pedido, para entender periodicidade, ticket médio, quantidade, distribuição pelo método de pagamento, motivo de compra e outras dimensões aplicáveis.
- Por região, para entender quais locais a Adventure Works possui uma parcela maior de mercado.

- Por clientes, para utilizar ações de e-mail marketing com os clientes que mais comprem e reaproximar aqueles que estão sem comprar há um tempo.

Após a etapa de entendimento dos requisitos iniciais, obteve-se algumas perguntas de negócio para nortear a modelagem:

- **Por Pedido**

- Qual é o ticket médio?

Métrica: Média do valor total por pedido.

- Qual é a distribuição dos pedidos por método de pagamento?

Métrica: Percentual de pedidos por tipo de pagamento.

- Quais são os principais motivos de compra informados pelos clientes?

Métrica: Frequência de cada motivo.

- Qual a sazonalidade das vendas?

Métrica: Quantidade de pedidos por período (dia, mês, trimestre).

- Qual é a média de itens por pedido?

Métrica: Média do número de itens por pedido.

- **Por Região**

- Quais regiões têm o maior faturamento?

Métrica: Receita total por região.

- Qual é a participação de mercado da Adventure Works em cada região?

Métrica: Receita da empresa em relação ao mercado total da região (market share).

- Quais produtos têm maior demanda em cada região?

Métrica: Quantidade de unidades vendidas por produto em cada região.

- **Por Clientes**

- Qual é o tempo médio de retenção de um cliente ativo?

Métrica: Tempo médio entre a primeira e a última compra de clientes ativos.

- Qual é a taxa de retorno de clientes?

Métrica: Percentual de clientes que realizaram mais de uma compra em um período.

- Como está o engajamento com campanhas de e-mail marketing?

Métrica: Taxa de abertura, taxa de cliques e conversões por campanha.

- Qual é o perfil demográfico dos clientes que mais compram?

Métrica: Segmentação por idade, gênero, renda ou outras características disponíveis.

Após o levantamento dessas perguntas, foi necessário identificar quais tabelas seriam usadas. A partir dos setores que deverão ser abordados, foram selecionadas:

Schema Sales:

- SalesOrderHeader
- SalesOrderDetail
- SalesReason
- SalesOrderHeaderSalesReason
- Customer
- Store
- CreditCard

Schema Person:

- Address
- StateProvince
- CountryRegion
- Person

Schema Production:

- Product
- ProductCategory
- ProductSubcategory

Com as tabelas identificadas, é necessário criar um diagrama conceitual que defina as tabelas fato e dimensões. A separação em tabelas fato e dimensões é importante para permitir a filtragem, o agrupamento e o resumo das ações de vendas. Essa estrutura facilita a organização dos dados, melhora o desempenho das análises e evita o uso desnecessário de informações, garantindo a padronização para uso futuro e a integração com outras tabelas.

Após a escolha das tabelas que serão utilizadas, dividiu-se o schema em fct_sales, que representa as informações das vendas com granularidade em produto por venda e as dimensões de território, motivo de venda, cartão de crédito, clientes e produtos. Também foi adicionada a tabela agregada de acordo com as especificações do problema.

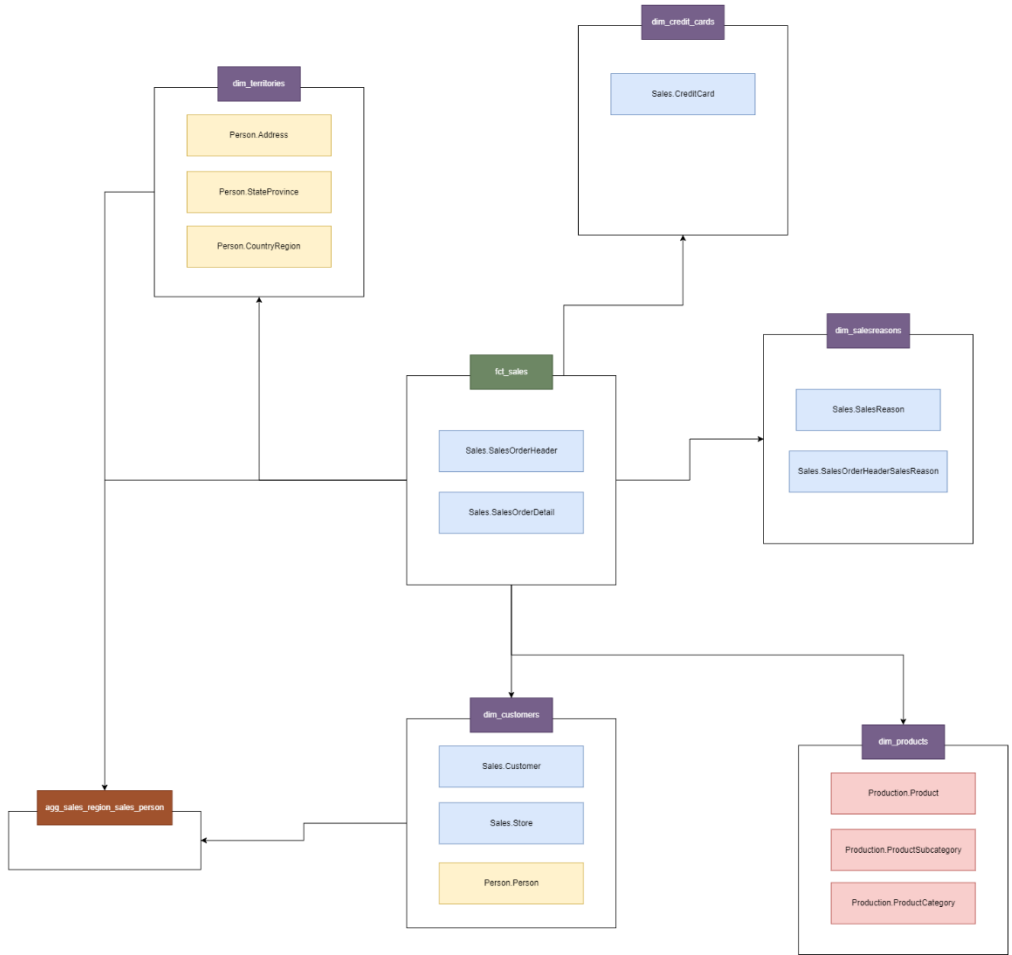


Diagrama Conceitual DW

4. Data Warehouse

Um **Data Warehouse (DW)** é um sistema de armazenamento de dados voltado para a análise e tomada de decisões estratégicas. Ele reúne informações de diferentes fontes, como sistemas transacionais e planilhas, organizando-as de forma estruturada e padronizada. Diferentemente de bancos de dados operacionais, que são projetados para lidar com transações do dia a dia, o DW é otimizado para consultas analíticas, permitindo explorar tendências, padrões e insights históricos.

Os dados armazenados em um DW são integrados, consolidados e organizados por temas relevantes para o negócio, como vendas ou clientes, e incluem informações históricas que possibilitam análises ao longo do tempo. Uma vez carregados, os dados permanecem estáticos, garantindo integridade e confiabilidade para relatórios e análises, sendo essencial para empresas que buscam transformar dados brutos em decisões informadas.

O DW para este projeto foi implementado na plataforma Google Cloud Platform (GCP) utilizando o BigQuery como solução de armazenamento.

4.1. Big Query

Os dados utilizados no projeto foram carregados diretamente a partir de um repositório no GitHub, sem passar pela etapa tradicional de extração. Para viabilizar a integração entre o GitHub, BigQuery e DBT Cloud, foram seguidos as etapas descritas abaixo.

Para iniciar o processo, foi criado um projeto no Google Cloud, que serviu como base para todas as configurações necessárias. Dentro do ambiente do Google Cloud, foi habilitada a API do BigQuery, permitindo o uso dos recursos de gerenciamento e análise de dados dessa ferramenta.

Em seguida, foi configurada uma conta de serviço no Google Cloud. Para isso, no Google Cloud Console, acessou-se o menu "IAM & Admin" e a seção de "Service Accounts". Uma nova conta de serviço foi criada, com um nome e descrição adequados, e foram atribuídas permissões específicas, como BigQuery Admin, garantindo o acesso necessário ao BigQuery. Após a criação da conta de serviço, foi gerada uma chave para ela. Para isso, acessou-se a conta criada, navegou-se até a aba "Keys" e foi selecionada a opção "Add Key > Create New Key". Um arquivo no formato JSON foi gerado e baixado, sendo armazenado em um local seguro para uso posterior.

Status da avaliação gratuita: R\$ 2.117,04 de crédito e 66 dias restantes. Ative sua conta completa para ter acesso ilimitado a todos os recursos.

Google Cloud

adventureworks-desafio

Pesquisar (/) recursos, documentos, produtos e serviços

IAM e administrador

IAM

PERMITIR

NEGAR

HISTÓRICO DE RECOMENDAÇÕES

Permissões do projeto "adventureworks-desafio"

Essas permissões afetam este projeto e todos os recursos dele. Saiba mais

VISUALIZAR POR PRINCIPAIS

VISUALIZAR POR PAPEIS

CONCEDER ACESSO

REMOVER ACESSO

Filtro

Insira o nome ou o valor da propriedade

<input type="checkbox"/>	Tipo	Principal	Nome	Papel
<input type="checkbox"/>		adventureworks@adventureworks-desafio.iam.gserviceaccount.com	adventureworks	Proprietário
<input type="checkbox"/>		dalviatattoo@gmail.com	Livia Nobre	Proprietário

Editar o acesso a "adventureworks-desafio"

Principal

adventureworks@adventureworks-desafio.iam.gserviceaccount.com

Projeto

adventureworks-desafio

Atribuir papéis

Os papéis são compostos de conjuntos de permissões e determinam o que o principal pode fazer com esse recurso. Saiba mais

Papel

Filtrar por função ou permissão

Acesso rápido

Em uso

Básico

Por produto ou serviço

Access Context Manager

Acesso VPC sem servidor

Papéis

Editor

Leitor

Navegador

Proprietário

GERENCIAR PAPEIS

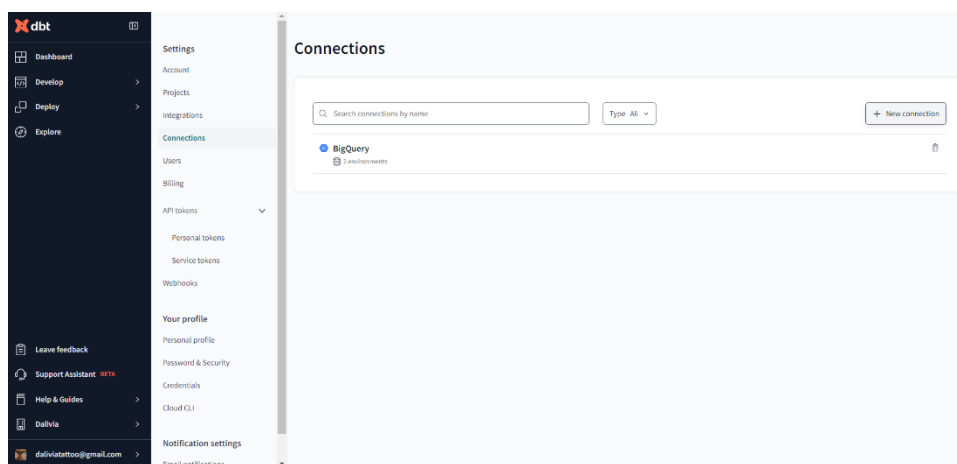
Configuração da Big Query

5. dbt Cloud

O DBT Cloud foi a ferramenta para a transformação de dados escolhida devido à sua capacidade de simplificar a modelagem de dados, utilizando o SQL como linguagem principal. Isso permite a criação de modelos de maneira intuitiva e declarativa, facilitando a definição de estruturas complexas e transformações específicas com uma sintaxe amplamente conhecida e acessível. Além disso, a ferramenta oferece integração facilitada com diversas plataformas, tornando-a uma solução robusta e eficiente para gestão e transformação de dados.

5.1. Conexão com Big Query

Para conectar o DBT Cloud ao BigQuery, foi necessário acessar a plataforma DBT Cloud por meio do link <https://cloud.getdbt.com> e realizar o login. No painel principal, navegou-se até as configurações da conta (Account Settings) e, em seguida, à seção de conexões (Connections) na barra lateral esquerda. Foi criada uma nova conexão clicando em "New Connection" e selecionando a opção "BigQuery" com o tipo de conexão desejado.



Connections no dbt cloud

Os campos obrigatórios da conexão foram preenchidos com as informações correspondentes. Foi atribuído um nome à conexão no campo "Connection Name", inserido o ID do projeto do Google Cloud no campo "Project ID", e especificado o nome do dataset do BigQuery a ser utilizado no campo "Dataset". Além disso, o arquivo JSON da chave de serviço criado anteriormente foi carregado no campo "Keyfile". Após salvar essas configurações, a conexão foi testada para garantir que estivesse funcionando corretamente.

Connection name
BigQuery

Settings

dbt Cloud will always access your connection from 52.3.77.232 , 3.214.191.138 , or 34.233.79.135 . Make sure to allow inbound traffic from these IPs in your firewall, and include it in any database grants.

Upload from file
[Upload a Service Account .JSON file](#)

Project ID

Job Execution Timeout Seconds

Use the job_execution_timeout_seconds configuration to set the number of seconds dbt should wait for queries to complete, after being submitted successfully.

Private Key ID

Private Key

Client email

Client ID

Preenchimento de credenciais para conexão no dbt Cloud

5.2. Conexão dbt e Github


No DBT Cloud, a conexão com o repositório GitHub foi configurada diretamente pelo perfil do usuário, acessando as configurações disponíveis no painel. Esse processo permitiu integrar o repositório ao ambiente de desenvolvimento, possibilitando o versionamento e a colaboração de forma prática. A interface do DBT Cloud facilita a configuração dessa integração, garantindo uma conexão rápida e eficiente com o repositório desejado.

dbt

Settings
Account
Projects
Integrations
Connections
Users
Billing
API tokens
Personal tokens
Service tokens
Webhooks
Your profile
Personal profile
Password & Security
Credentials
Cloud CLI
Notification settings
Email notifications
Slack notifications

User profile

Personal information [Edit](#)





First name

Last name

Email
daliviatattoo@gmail.com
We'll send a verification email to your new email address. You must verify in order to complete this change.

Linked accounts

>  GitHub **LiviaNobre** [X](#)

>  GitLab [Link](#)

Experimental features [Experimental features can be discontinued without notice.](#)

Painel de conexão no dbt

5.3. Transformações dos dados

Para a construção do pipeline e extração de insights, foi adotada a abordagem ELT (Extração, Carregamento e Transformação). Nesse processo, os dados são inicialmente carregados no Data Warehouse (DW) — neste caso, o BigQuery — para, em seguida, passarem pelas transformações realizadas no DBT. Após as alterações, os dados transformados são devolvidos ao DW, permitindo que os insights sejam obtidos de forma estruturada e otimizada.

5.3.1. dbt_project.yml

Os arquivos **.yml** são usados para definir fontes de dados, schemas e organizar fluxos de execução. Para iniciar o projeto, as configurações básicas devem ser colocadas no documento chamado `dbt_project.yml`, arquivo que contém configurações fundamentais que permitem ao DBT localizar a base de dados a ser utilizada. Nesse arquivo, são definidas informações como o nome do banco de dados e as configurações para a materialização dos modelos, que podem ser criados como visualizações ou transformados em tabelas persistentes.

5.3.2. Staging

A camada de *staging* serve como uma área temporária onde os dados extraídos dos sistemas de origem são armazenados antes de serem transformados. Em vez de acessar os dados diretamente da fonte, o DBT utiliza essa camada como intermediária, possibilitando o tratamento e a entrega de dados de forma mais eficiente e controlada.

O processo de configuração começa com a definição das tabelas no arquivo **source.yml**, onde são inseridas as credenciais que permitem ao dbt identificar as fontes de dados. Após isso, cada tabela é detalhada com informações como nome, descrição e testes específicos — como a verificação de valores únicos ou nulos — que podem ser aplicados à tabela inteira ou a colunas específicas.

Abaixo está um exemplo de como o arquivo `source.yml` fica após o preenchimento. Ao finalizar a configuração de cada tabela, a funcionalidade *Generate Model* pode ser utilizada para carregar automaticamente o modelo no formato `sql` no dbt.

models / staging / source.yml

```
1  version: 2
2
3  sources:
4    - name: sap_adw
5      description: Fonte do SAP do Adventure Works
6      schema: sap_adw
7      tables:
8        - name: customer
9          description: tabela de clientes
10
11         columns:
12           - name: customerid
13             description: Id do cliente. Chave primária (PK) da tabela
14             data_type: int64
15             tests:
16               - unique
17
```

Preenchimento do arquivo source.yml

Com o modelo gerado, inicia-se o processo das primeiras transformações. A camada de *staging* é destinada exclusivamente às transformações que não alteram a estrutura original das tabelas. Nessa etapa, foram realizadas modificações como a renomeação de colunas e a alteração de tipos de variáveis, mantendo a integridade da tabela.

Todas as tabelas previamente selecionadas foram carregadas para a área de *staging*, onde passaram pelas primeiras alterações necessárias, garantindo que os dados estejam padronizados e prontos para etapas posteriores de transformação.

staging

source.yml

- stg_sap_adw__address.sql
- stg_sap_adw__countryregion.sql
- stg_sap_adw__creditcard.sql
- stg_sap_adw__customer.sql
- stg_sap_adw__person.sql
- stg_sap_adw__product.sql
- stg_sap_adw__productcategory.sql
- stg_sap_adw__productsubcategory.sql
- stg_sap_adw__salesorderdetail.sql
- stg_sap_adw__salesorderheader.sql
- stg_sap_adw__salesorderheadersalesreason.sql
- stg_sap_adw__salesreason.sql
- stg_sap_adw__stateprovince.sql
- stg_sap_adw__store.sql

Tabelas na camada staging no dbt

stg_sap_adw__address	☆	:
stg_sap_adw__countryregion	☆	:
stg_sap_adw__creditcard	☆	:
stg_sap_adw__customer	☆	:
stg_sap_adw__person	☆	:
stg_sap_adw__product	☆	:
stg_sap_adw__productcategory	☆	:
stg_sap_adw__productsubcategory	☆	:
stg_sap_adw__salesorderdetail	☆	:
stg_sap_adw__salesorderheader	☆	:
stg_sap_adw__salesorderheadersalesreason	☆	:
stg_sap_adw__salesreason	☆	:
stg_sap_adw__stateprovince	☆	:
stg_sap_adw__store	☆	:

Tabela na camada staging na BQ

5.3.3. Marts

Na camada *Marts*, são realizadas as transformações mais avançadas, envolvendo junções e agregações dos dados. É nessa etapa que são criadas as tabelas fato e dimensões, essenciais para a construção de um modelo analítico robusto. Para isso, foram utilizados *joins* para integrar as tabelas com as informações necessárias.

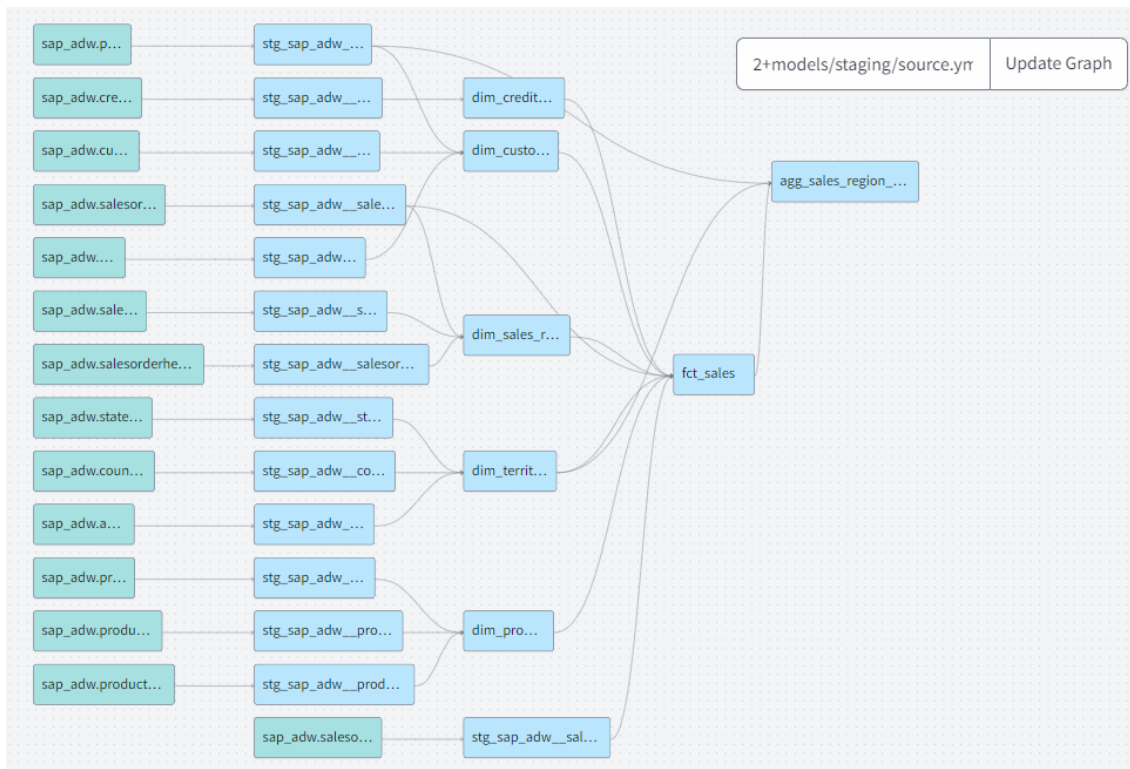
Além disso, para cada tabela da camada *Marts*, foi desenvolvido um arquivo *.yml* destinado à execução de testes, assegurando a integridade e consistência das informações geradas. Foram criadas cinco tabelas dimensão a partir das tabelas *staging*, além de uma dimensão adicional, elaborada com base em um código padrão, para a construção da *dim_date*. Esta tabela desempenha um papel fundamental como suporte para futuros modelos que utilizem dados temporais.

Uma outra tabela criada foi a agregada de vendas por região e vendedor. Nesta tabela foram adicionadas algumas métricas que podem ser usadas na construção do dashboard.

```
select
  fct_sales.salespersonid
  , region.stateprovince_name
  , count(distinct fct_sales.salesorderid) as total_orders
  , sum(fct_sales.orderqty) as total_products_qty -- total de produtos comprados
  , sum(fct_sales.totaldue) as total_sales_value -- receita total
  , sum(fct_sales.subtotal) as total_subtotal
  , avg(fct_sales.taxamt) as total_tax
  , avg(fct_sales.freight) as total_freight
  , avg(fct_sales.totaldue) as average_order_value
  , count(distinct fct_sales.salesorderid) as total_qty_sales
  , count(distinct fct_sales.salespersonid) as total_salespersons
from fct_sales
left join person on person.businessentityid = fct_sales.salespersonid
left join region on region.addressid = fct_sales.shiptoaddressid
group by salespersonid, stateprovince_name
```

Código para a criação da tabela agregada de vendas por região e vendedor

Os códigos referentes às transformações das *Marts*, *Staging* e arquivos *.yml* encontram-se disponíveis no repositório do projeto, permitindo a consulta e replicação do trabalho.



Lineage do projeto da AW no dbt

5.4. Ambiente de Produção

Para garantir que o processo de transformação de dados seja realizado de forma automatizada e consistente, foi configurado um ambiente de produção no dbt Cloud, direcionando as transformações para um schema de destino específico. Com isso, a produção de dados transformados se tornou parte de um fluxo contínuo e controlado. Além disso, foi criada uma rotina de jobs no dbt Cloud, que executa os comandos de transformação (como dbt run, dbt test, e dbt build) em intervalos programados. Esses jobs asseguram que as transformações sejam realizadas de forma regular e sem a necessidade de intervenção manual. O dbt Cloud também oferece logs detalhados, que ajudam a monitorar o andamento do processo e a diagnosticar qualquer possível problema, garantindo um pipeline de dados bem gerido e eficiente.

Esta etapa será mais explorada na orquestração do Airflow, já que é parte integrada do passo a passo para uma conexão correta.

6. Orquestração com Apache Airflow

A configuração do Airflow para rodar Jobs no DBT Cloud foi realizada a partir do tutorial disponibilizado pelo próprio dbt, que foi obtido através deste link: <https://docs.getdbt.com/guides/airflow-and-dbt-cloud?step=1>*. No repositório do desafio também há um tutorial mais detalhado com imagens e explicações que esmiuça os passos.

1. Instalação do CLI Astro

O primeiro passo foi a instalação do CLI Astro, ferramenta necessária para gerenciar o ambiente do Airflow. O download foi feito a partir do site oficial do Astro CLI ou diretamente do repositório no GitHub. Escolheu-se a versão compatível com o sistema operacional, seja ela "amd64" ou "arm64".

Após o download, o arquivo foi renomeado para astro.exe e movido para uma nova pasta chamada astro, localizada no diretório do projeto. O caminho dessa pasta foi copiado para uso posterior na configuração.

No Windows, o próximo passo foi adicionar o caminho da pasta astro à variável de ambiente PATH. Para isso, foi necessário acessar a opção "Editar Variáveis de Ambiente" no menu iniciar, selecionar a variável PATH e inserir o caminho da pasta astro. Após salvar as alterações, abriu-se o PowerShell para verificar a instalação com o comando astro version. Com isso, foi confirmada a instalação do CLI Astro.

2. Instalação do Docker

A instalação do Docker foi realizada baixando o instalador para Windows através do site oficial. Foi selecionada a versão x86_64, padrão para a maioria dos computadores. O processo de instalação foi simples, sem a necessidade de configurações adicionais.

3. Clonagem do Repositório DBT Cloud

Com o Docker e o CLI Astro configurados, o próximo passo foi clonar o repositório do DBT Cloud para o projeto. Isso foi feito via PowerShell com o comando git clone. Para quem não possui uma chave SSH configurada, também é possível baixar o repositório em formato ZIP diretamente do site do GitHub.

Após o download, foram copiados para o diretório do projeto os seguintes arquivos e pastas do repositório clonado: .astro, dags, .dockerignore, .python-

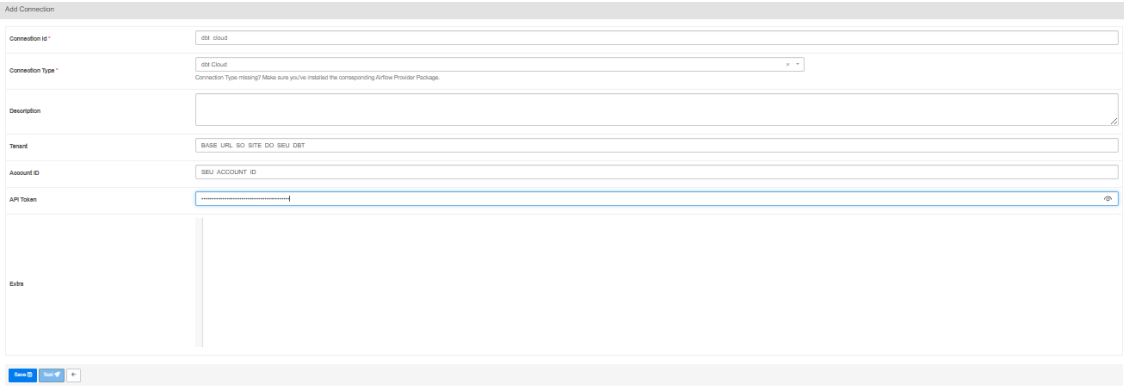
version, Dockerfile, LICENSE, packages e requirements.txt. Esses arquivos são os necessários para o funcionamento do Airflow integrado ao DBT.

Com os arquivos no lugar, o Airflow foi iniciado utilizando o comando `astro dev start` no PowerShell. Esse comando realizou a configuração necessária para rodar o Airflow, que ficou acessível no navegador pelo endereço <http://localhost:8080>. As credenciais padrão do Airflow foram:

- **Username:** admin
- **Password:** admin

4. Conexão do DBT Cloud com o Airflow

A integração entre o DBT Cloud e o Airflow exigiu a criação de um Service Token no DBT Cloud. Esse token foi gerado no menu "Account Settings > Token API > Service Token". Foi ressaltado que o token deve ser do tipo *Service Token*, pois *Personal Tokens* não permitem a execução de *Jobs*. Caso o período de teste gratuito do DBT Cloud tenha expirado, é recomendada a criação de uma nova conta.



The screenshot shows the 'Add Connection' form in the Airflow web interface. The form is titled 'Add Connection' and contains the following fields:

- Connection ID:** A text input field with the value 'dbt-cloud'.
- Connection Type:** A dropdown menu with 'dbt Cloud' selected. Below the dropdown, a small text note reads: 'Connection Type missing? Make sure you've installed the corresponding Airflow Provider Package.'
- Description:** A text input field.
- Tenant:** A text input field with the value 'BASE URL DO SITE DO SEU DBT'.
- Account ID:** A text input field with the value 'SEU ACCOUNT ID'.
- API Token:** A text input field with a password icon on the right.
- Extra:** A large text area for additional configuration.

At the bottom of the form, there are buttons for 'Save', 'Cancel', and 'Test'.

Configuração connection dbt

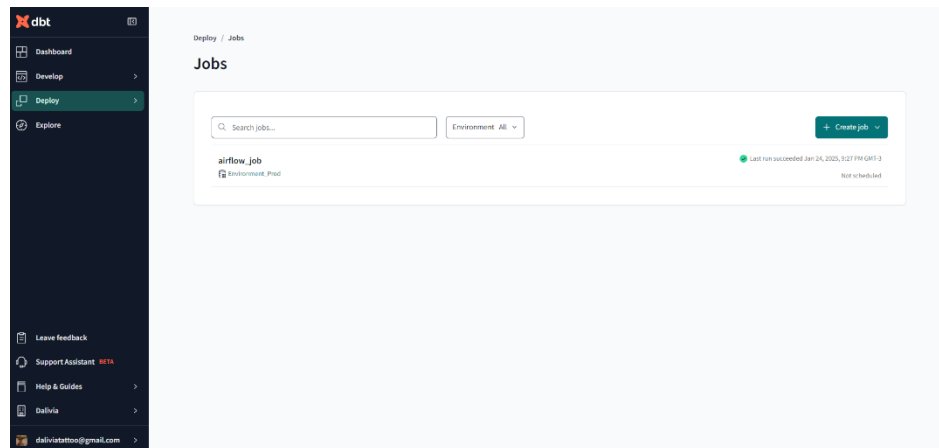
Com o token gerado, foi necessário configurá-lo no Airflow. No menu "Admin > Connections", foi criada uma nova conexão com as seguintes informações:

- **Connection ID:** nome escolhido para identificar a conexão.
- **Connection Type:** selecionado como `dbt_cloud`.
- **Tenant:** base URL do projeto no DBT Cloud (exemplo: `getdbt.com`).
- **Account ID:** ID da conta no DBT Cloud.
- **API Token:** o token gerado no DBT Cloud.

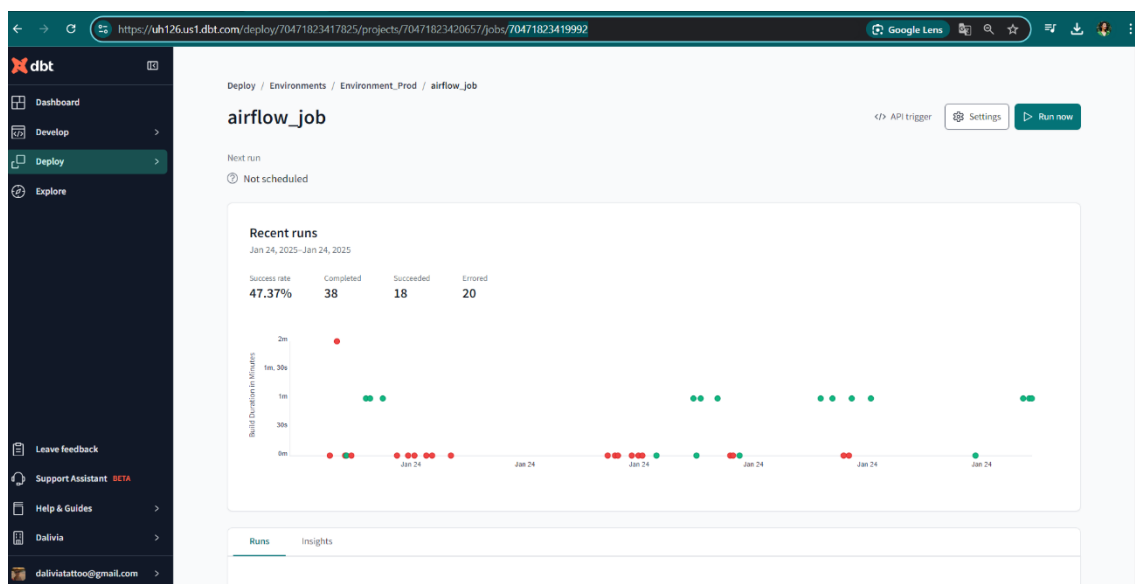
Após preencher esses dados, a conexão foi salva.

5. Criação e Execução de um Job no DBT Cloud

No DBT Cloud, foi criado um ambiente de produção através do menu "Deploy > Environment". Nesse ambiente, foi configurada uma branch específica (opcional) ou definida a branch padrão como main. No menu "Deploy > Job", foi configurado um *Job* com os comandos dbt test e dbt run. O ID do Job foi salvo, pois seria utilizado posteriormente na configuração da DAG no Airflow.



Job no dbt



6. Criação da DAG no Airflow

Para configurar a DAG (Directed Acyclic Graph), foi criado um arquivo Python chamado `dbt_cloud_run_job.py` na pasta `dags` do projeto. O código da DAG foi estruturado para rodar diariamente às 6h da manhã, utilizando o operador `DbtCloudRunJobOperator` para executar o Job configurado no DBT Cloud. Abaixo está o exemplo de código usado na DAG:

```
from datetime import datetime

from airflow.models import DAG
from airflow.providers.dbt.cloud.operators.dbt import DbtCloudRunJobOperator

### Update these ids to match your account ###
DBT_CLOUD_CONN_ID = <NOME_DA_MINHA_CONEXAO_DBT>
ACCOUNT_ID = <ACCOUNT_ID_DAMINHA_CONTA_DBT>
JOB_ID = <JOB_ID_CRIADO_PARA_ESTES_DESAFIO>

with DAG(
    # usar conexão criada no airflow
    dag_id="dbt_cloud_run_job",
    default_args={"dbt_cloud_conn_id": DBT_CLOUD_CONN_ID, "account_id": ACCOUNT_ID},
    start_date=datetime(2021, 1, 1),
    # rodar às 6h diariamente
    schedule_interval='0 6 * * *',
    catchup=False,
) as dag:

    trigger_dbt_cloud_job_run = DbtCloudRunJobOperator(
        task_id="trigger_dbt_cloud_job_run",
        job_id=JOB_ID,
        check_interval=10,
        timeout=300,
        retry_from_failure=True,
    )

    trigger_dbt_cloud_job_run
```

Código de criação da DAG

7. Verificação da Execução da DAG

Para garantir que o Airflow estava funcionando corretamente e executando a DAG conforme esperado, a DAG foi agendada para rodar automaticamente às 6h da manhã. Após rodar o processo, os logs confirmaram que a execução do Job estava sendo realizada com sucesso, como pode ser visto nas imagens abaixo, que mostram o histórico de execuções da DAG.

DAGs

All 1Active 1Paused 0

Running 0Failed 0

Filter DAGs by tag

Search DAGs

Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
dbt_cloud_run_job	airflow	4	Schedule: At 06:00	2025-01-24, 21:32:17	2025-01-24, 06:00:00	1		

Showing 1-1 of 1 DAGs

Astronomer Runtime 12.1.0 based on Airflow 2.10.1+astro.1
Git Version: release.7a11fe6438b5ea81c75c4e5a356a6c23ab18404

Prova Schedule

Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

Astronomer

21:40 UTC

AU

DAG: dbt_cloud_run_job

Schedule: 0 6 * * *Next Run ID: 2025-01-24, 06:00:00 UTC

24/01/2025 21:40:14All Run TypesAll Run StatesClear FiltersAuto-refresh 25

trigger_dbt_cloud_job_run

dbt_cloud_run_job

Details

Graph

Gantt

Code

Event Log

Run Duration

Task Duration

Calendar

DAG Runs Summary

Total Runs Displayed	8
Total success	4
Total failed	4
First Run Start	2025-01-24, 18:16:02 UTC
Last Run Start	2025-01-24, 21:32:17 UTC
Max Run Duration	00:01:21
Mean Run Duration	00:00:37
Min Run Duration	00:00:01

DAG rodando 3 vezes

List Dag Run

Search

Actions

Record Count: 4

	State	Dag Id	Logical Date	Run Id	Run Type	Queued At	Start Date	End Date	Note	External Trigger	Conf	Duration
	success	dbt_cloud_run_job	2025-01-24, 21:32:17	manual__2025-01-24T21:32:17.415454+00:00	manual	2025-01-24, 21:32:17	2025-01-24, 21:32:17	2025-01-24, 21:33:26	True			1M:8s
	success	dbt_cloud_run_job	2025-01-24, 21:30:04	manual__2025-01-24T21:30:04.598064+00:00	manual	2025-01-24, 21:30:04	2025-01-24, 21:30:04	2025-01-24, 21:31:02	True			57s
	success	dbt_cloud_run_job	2025-01-24, 21:28:26	manual__2025-01-24T21:28:26.728806+00:00	manual	2025-01-24, 21:28:26	2025-01-24, 21:28:27	2025-01-24, 21:29:49	True			1M:21s
	success	dbt_cloud_run_job	2025-01-24, 18:38:27	manual__2025-01-24T18:38:27.217168+00:00	manual	2025-01-24, 18:38:27	2025-01-24, 18:38:28	2025-01-24, 18:39:50	True			1M:21s

Astronomer Runtime 12.1.0 based on Airflow 2.10.1+astro.1

Lista de sucessos da DAG

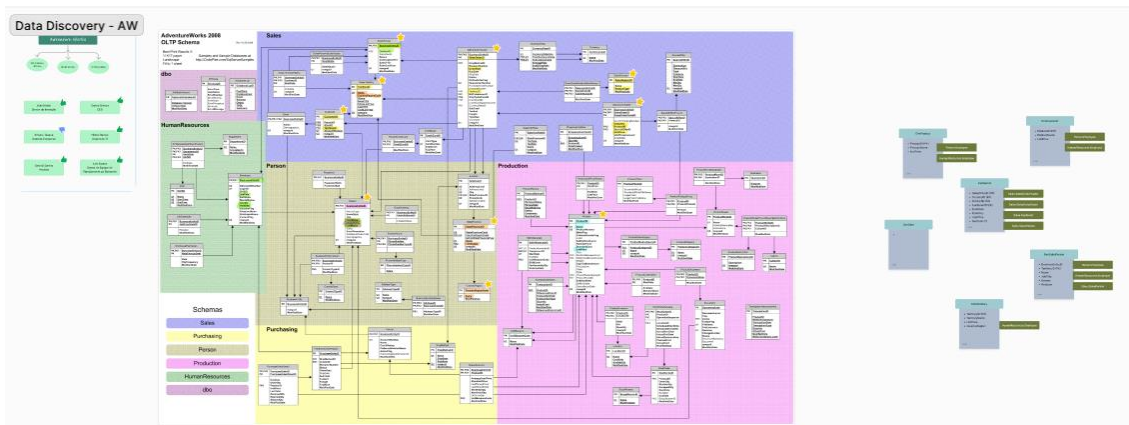
8. Apresentação dos dados

8.1. Mockup

Após a etapa de transformação dos dados, foi necessário planejar a visualização de dados, considerando como essas informações poderiam agregar valor ao negócio. O primeiro passo foi definir os elementos essenciais que deveriam estar presentes no dashboard e no relatório destinado ao CEO.

Com base em perguntas de negócio previamente estabelecidas, o dashboard foi estruturado para apresentar, inicialmente, uma visão geral da empresa, com informações mais globais. Nas páginas seguintes, o foco se concentrou em áreas específicas que demandavam análises mais aprofundadas, como pedido, região e cliente. Essa organização permitiu um equilíbrio entre uma visão macro e o detalhamento necessário para tomadas de decisão estratégicas.

Para a escolha da paleta de cores e a criação do mockup, foram utilizadas as ferramentas **Figma** e **FigJam**. O FigJam foi empregado em um processo inicial de *data discovery*, proporcionando uma visualização ampla das demandas da empresa e auxiliando na seleção das informações que deveriam ser destacadas no dashboard.



Uso do Figjam para a elaboração de um Data Discovery

Já o Figma foi utilizado para a criação dos mockups tanto do dashboard quanto do relatório. As funcionalidades intuitivas e direcionadas da ferramenta possibilitaram a experimentação com diferentes paletas de cores e layouts, resultando em um design final que atende aos objetivos de clareza e funcionalidade.

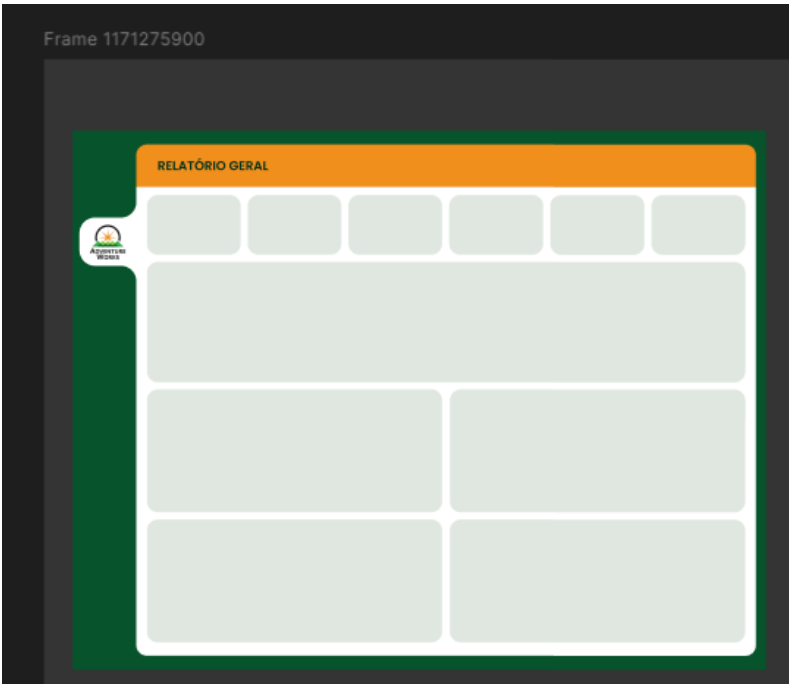
Abaixo estão os modelos finais, com a paleta de cores da empresa, a logo, além dos espaçamentos para os campos que será preenchido no Power Bi. Cada página teve seu layout adaptado para o que seria mostrado nela. Também em anexo está o mockup para o relatório do CEO, que seguiu a mesma estética do dashboard e a paleta de cores encontrada através de uma extensão do próprio Figma.



Mockup da primeira página do dashboard



Capa e página de cliente do dashboard



Mockup do relatório para o CEO

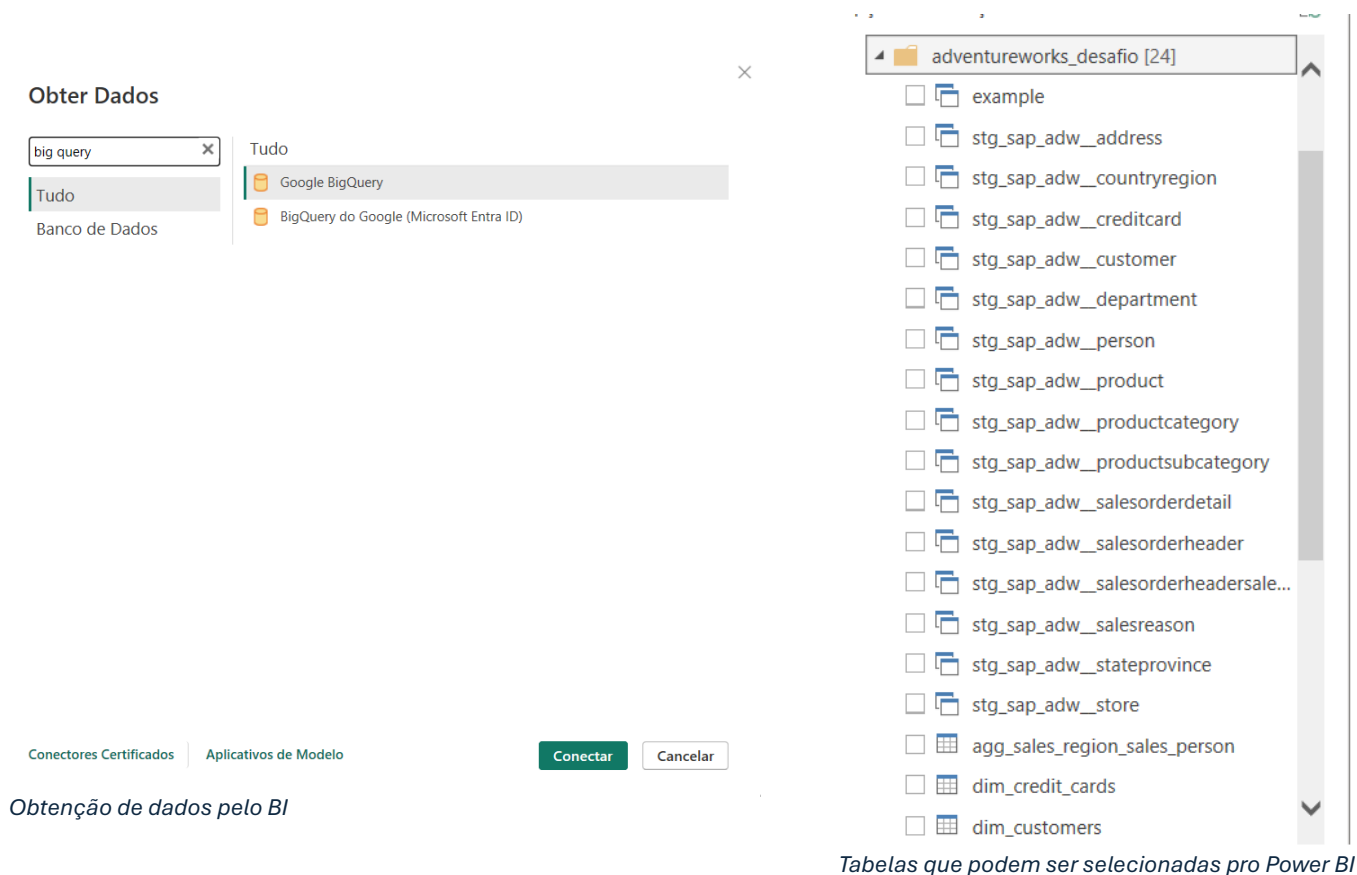


Paleta obtida pela logo da marca

8.2. Conexão do Power BI com a BQ

A ferramenta escolhida para a visualização dos dados foi o Power Bi, uma plataforma com visualizações interativas, que possibilita uma rápida atualização dos dados além de seu uso ser bastante intuitivo.

Para a criação do dashboard e do relatório, deve-se acessar os dados armazenados no BigQuery após a transformação. O processo se inicia ao selecionar a opção "Obter Dados" e, em seguida, localizar a integração com o Google Cloud. Com as contas devidamente conectadas, o Power BI reconhece automaticamente as tabelas disponíveis no BigQuery, permitindo a seleção das tabelas relevantes para a análise e construção das visualizações.



Ao selecionar as tabelas que serão usadas, o Power Bi permite dois tipos de conexões diferentes com a Big Query: Importar ou DirectQuery. O DirectQuery puxa os dados diretamente da BQ, podendo ser atualizados em tempo real. Já ao importar os dados, é feita uma cópia dos dados até o momento do carregamento, possibilitando mais funcionalidades na hora de transformar os dados no próprio BI.

Configurações de conexão

Você pode escolher como se conectar a esta fonte de dados. A importação permite que você traga uma cópia dos dados para o Power BI. O DirectQuery se conectará a esta fonte de dados em tempo real.

☒ Importar

☐ DirectQuery

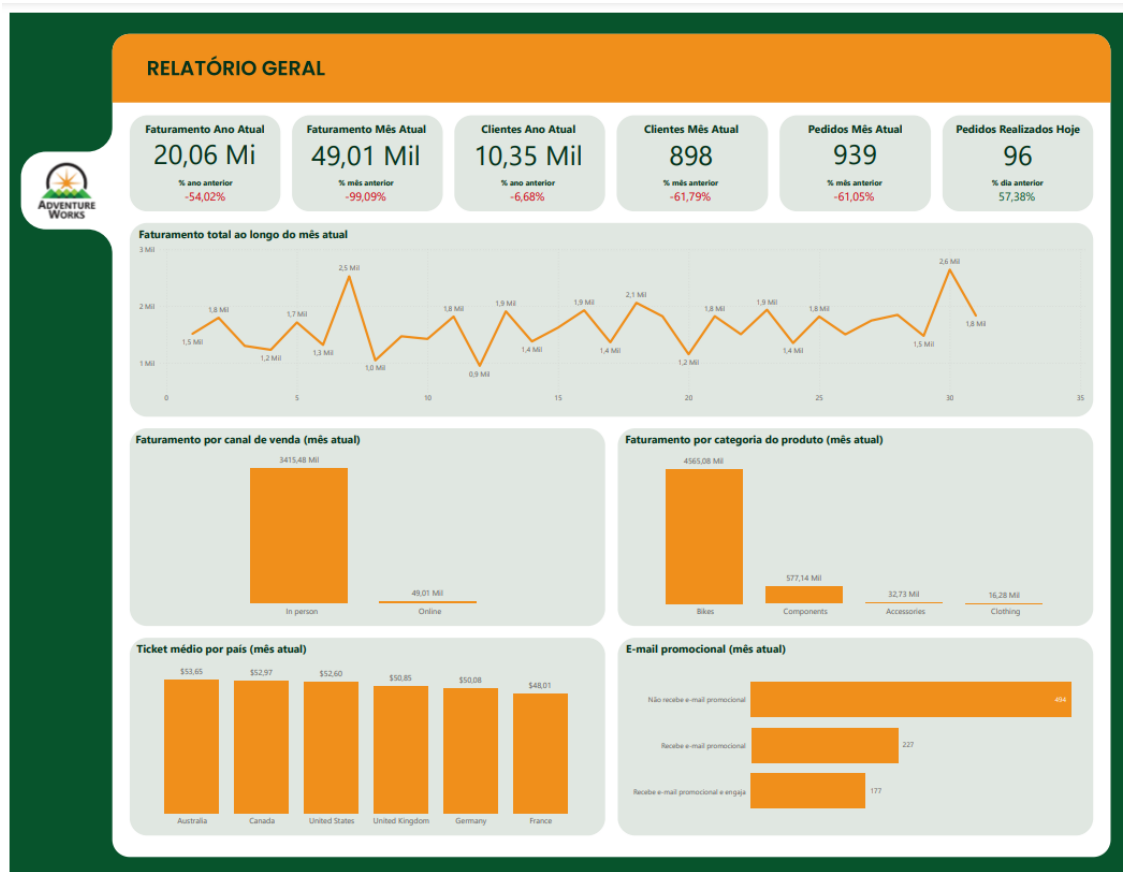
[Saiba mais sobre o DirectQuery](#)

OK

Cancelar

Tipos de conexões com as fontes de dados

8.3. Relatório Geral para o CEO



Relatório para o CEO

O relatório mostra alguns indicadores que possuem um comparativo com o período anterior, assim o CEO pode ter um acompanhamento do comportamento das métricas com o passar do tempo, já que os dados são carregados diariamente.

O dashboard foi elaborado para oferecer uma visão abrangente e detalhada dos principais indicadores de desempenho, com foco em métricas financeiras, operacionais e de marketing.

8.3.1. Métricas

- Faturamento Anual e Mensal Atual

O faturamento anual fornece uma visão consolidada do desempenho financeiro ao longo do ano, enquanto o faturamento mensal permite análises de curto prazo. Ambos os indicadores são essenciais para entender a evolução das receitas e identificar sazonalidades.

Para cada período (ano e mês), apresenta-se a porcentagem de variação em relação ao período equivalente anterior (ano anterior ou mês anterior), fornecendo contexto para crescimento ou declínio.

- Novos Clientes no Ano e no Mês Atual

Monitorar a quantidade de novos clientes permite avaliar a eficácia das estratégias de aquisição e a expansão da base de consumidores. Comparar os valores do mês e ano atuais com os mesmos períodos anteriores ajuda a entender o impacto de campanhas ou ações específicas.

- Pedidos do Mês Atual e Pedidos Realizados Hoje

A métrica de pedidos mensais avalia o volume de transações realizadas, oferecendo insights sobre a atividade comercial no mês. Já os pedidos diários, juntamente com a variação percentual em relação ao dia anterior, permitem monitorar em tempo real o desempenho e identificar mudanças abruptas.

8.3.2. Gráficos

Os gráficos foram desenvolvidos para oferecer uma análise mais detalhada do mês atual, com o objetivo de fornecer ao CEO uma compreensão aprofundada dos resultados recentes. A escolha dos gráficos foi guiada pela necessidade de insights acionáveis e pela facilidade de interpretação:

- Gráfico de Faturamento ao Longo do Mês Atual

Exibe o faturamento diário ao longo do mês, permitindo identificar os dias e períodos de maior venda. Essa análise é crucial para entender como eventos,

promoções ou sazonalidades afetam o desempenho diário e ajustar estratégias em tempo real.

- Gráfico de Faturamento por Canal de Venda (Mês Atual)

Mostra a divisão do faturamento entre os canais presencial e online. Isso permite avaliar o desempenho de cada canal, identificar oportunidades de melhoria e priorizar investimentos.

- Gráfico de Faturamento por Categoria de Produto (Mês Atual)

Detalha o faturamento por categorias como bikes, components, accessories e clothing. Essa visualização auxilia na identificação das categorias mais lucrativas, guiando decisões de estoque, precificação e campanhas específicas.

- Gráfico de Ticket Médio por País (Último Mês)

Mostra o ticket médio por país, comparando mercados como Austrália, Canadá, Estados Unidos, Reino Unido, Alemanha e França. Esse gráfico ajuda a identificar regiões com maior poder aquisitivo ou demanda, facilitando a criação de estratégias regionais de precificação e marketing.

- Gráfico de E-mails Promocionais (Mês Atual)

Mede a eficácia das campanhas de e-mail marketing. A segmentação por comportamento (recebeu ou engajou) ajuda a avaliar o impacto das mensagens e identificar oportunidades de personalização.

A seleção criteriosa das métricas e gráficos permite ao CEO obter insights claros e direcionados para a tomada de decisão. As métricas financeiras, como faturamento e pedidos, fornecem uma visão do desempenho geral, enquanto os gráficos detalhados ajudam a entender o impacto de fatores específicos, como canal de venda, categorias de produtos e campanhas de marketing. Essa abordagem combina análise macro e micro, garantindo que tanto as tendências globais quanto os detalhes operacionais sejam considerados na gestão estratégica da empresa.

8.4. Dashboard

8.4.1. Visão Geral



Aba de Visão Geral do dashboard

Essa página de "Visão Geral" é essencialmente um painel executivo que fornece um panorama amplo da performance da empresa, funcionando como um ponto de partida para análises mais profundas. Aqui estão mais algumas informações e interpretações que podem ser extraídas:

→ Métricas

Faturamento Total (\$109,85M):

Este indicador reflete o sucesso geral da empresa ao longo do tempo. Pode ser usado como base para calcular indicadores como **Crescimento Anual Composto (CAGR)** ou para comparar o desempenho ao longo dos anos, considerando fatores como inflação, expansão de mercado ou novas aquisições.

Quantidade Total de Pedidos (31,47 Mil):

Indica o volume operacional, mostrando o quão ativa a empresa está no mercado. A relação entre faturamento total e quantidade de pedidos pode ajudar a calcular o **Ticket Médio Global**, o que pode revelar tendências no comportamento do consumidor.

Clientes Ativos (19,12 Mil):

Fornecer uma visão da base de clientes engajada. Essa métrica é importante para analisar a **retenção de clientes** e o impacto das estratégias de fidelização ou aquisição.

Lojas (702):

Esse número pode ser cruzado com o faturamento para avaliar o **faturamento médio por loja**. Também pode ser analisado para entender a relação entre expansão física e crescimento do faturamento.

→ Gráficos

Faturamento ao longo do tempo:

Picos visíveis em períodos específicos podem ser associados a sazonalidades (ex.: Black Friday, Natal) ou ações promocionais. Quedas abruptas podem indicar crises econômicas, mudanças de mercado ou até problemas internos, como rupturas de estoque. Esta visualização permite um acompanhamento constante do faturamento e a materialização do desempenho da empresa em diferentes momentos.

Faturamento por país:

A distribuição geográfica é um indicativo claro de onde está concentrada a força de mercado da empresa. Países com baixo faturamento podem ser explorados para expansão, enquanto países com alta receita podem ser usados como benchmarks para replicar estratégias em outros locais.

Os Estados Unidos detêm o maior número de loja e maior faturamento, o que pode guiar dois diferentes caminhos na visão do marketing: A implementação de novas estratégias para retenção do público em lugares que já estão com presença consolidada ou direcionar as ações de captura de cliente em lugares que não tem um bom desempenho.

Neste gráfico temos uma visão mais ampla, mas nas próximas páginas é possível desmembrar ainda mais as informações sobre as regiões.

Ticket médio por país:

Essa métrica é fundamental para identificar o comportamento de consumo em diferentes mercados. Países com maior ticket médio indicam uma disposição maior para gastos e podem justificar estratégias premium. Já mercados com ticket baixo podem demandar estratégias promocionais ou reposicionamento de preços.

Aqui estão algumas perguntas de negócio que podem ser respondidas através das visualizações:

- Como o faturamento total tem evoluído ao longo do tempo? Há sazonalidade ou picos em meses específicos?
- Quais mercados (países) contribuem mais para o faturamento total?
- O ticket médio varia significativamente entre países? Quais estratégias podem ser aplicadas para aumentar o ticket médio em países com valores menores?
- Como os diferentes canais de venda ou categorias de produtos impactam o faturamento total e a distribuição por países?

Esta página possui os seguintes filtros:

Canal de venda

Qual canal está gerando mais receita? Há diferenças no comportamento de consumidores online versus presencial? Se houver uma dependência excessiva de um canal, pode ser uma oportunidade para diversificar.

Categoria/Subcategoria de produto

Quais produtos lideram o faturamento em cada país ou canal? A performance de subcategorias pode indicar onde concentrar esforços de marketing ou expandir linhas de produtos.

Data

O filtro temporal permite análises de tendências sazonais e a medição do impacto de campanhas em diferentes períodos.

Insights da aba de Visão Geral:

- **Tendências no faturamento:** O gráfico de linha permite identificar períodos de maior faturamento, podendo ser associados a campanhas, sazonalidades ou lançamentos de produtos.
- **País com maior relevância:** Os Estados Unidos são o mercado mais importante (faturamento de \$63M), indicando a necessidade de manter estratégias agressivas nesse mercado.
- **Oportunidade em outros países:** Países como Alemanha (\$5M) ou Reino Unido (\$8M) apresentam um faturamento menor. Podem ser exploradas estratégias para aumentar a presença nesses mercados.
- **Diferenças no ticket médio:** O ticket médio dos Estados Unidos (\$5,2 Mil) é significativamente maior que em países como Austrália (\$1,6 Mil).

Estratégias como pacotes ou descontos progressivos podem ser testadas em países com menor ticket médio.

- **Correlação entre clientes ativos e faturamento:** A análise do impacto de clientes ativos versus lojas pode revelar a eficiência de expansão geográfica no suporte ao crescimento da base de clientes e ao aumento das receitas.

8.4.2. Pedido



Aba de pedido

Essa página foca na performance de vendas por categorias específicas de produtos, métodos de pagamento, e fatores que motivam as compras. Ela também explora dados como ticket médio, valor médio dos produtos e lifetime value.

→ Métricas

Ticket Médio (\$3,49 Mil):

Reflete o valor médio por pedido e permite identificar mudanças no comportamento de compra ao longo do tempo.

Valor Médio dos Produtos (\$535,86):

Fornecer uma visão detalhada do posicionamento de preços e como os produtos estão sendo comercializados.

Média de Produtos por Pedido (8,74):

Indica o volume médio de itens comprados por pedido, fundamental para entender hábitos de consumo.

Lifetime Value (LTV - \$122,77 Mil):

Avalia a receita média gerada por cliente ao longo de seu relacionamento com a empresa. Esta é uma métrica muito importante para decisões estratégicas de retenção e aquisição de clientes.

→ Gráficos

Faturamento por categoria de produto:

"Bikes" domina claramente (\$95M), seguido por componentes (\$12M) e outras categorias menores. Revela a importância estratégica de produtos "carros-chefe" e onde há espaço para expandir categorias menos lucrativas.

Quantidade de produtos vendidos x Ticket Médio:

Demonstra o volume de vendas por ano versus o ticket médio. Nota-se um declínio no ticket médio enquanto o volume de produtos vendidos cresce, indicando que promoções ou mudanças de mix de produtos podem estar influenciando o valor percebido.

Faturamento dos principais métodos de pagamento:

"ColonialVoice" lidera (\$32M), seguido por "Vista" (\$28M) e "SuperiorCard" (\$25M). Importante para identificar parcerias estratégicas e preferências regionais de pagamento.

Principais motivos de compra:

Preço é o maior motivador (14,7M compras), seguido por promoções e qualidade do fabricante. Isso ajuda a alinhar estratégias de marketing e a comunicar valor de maneira mais eficiente.

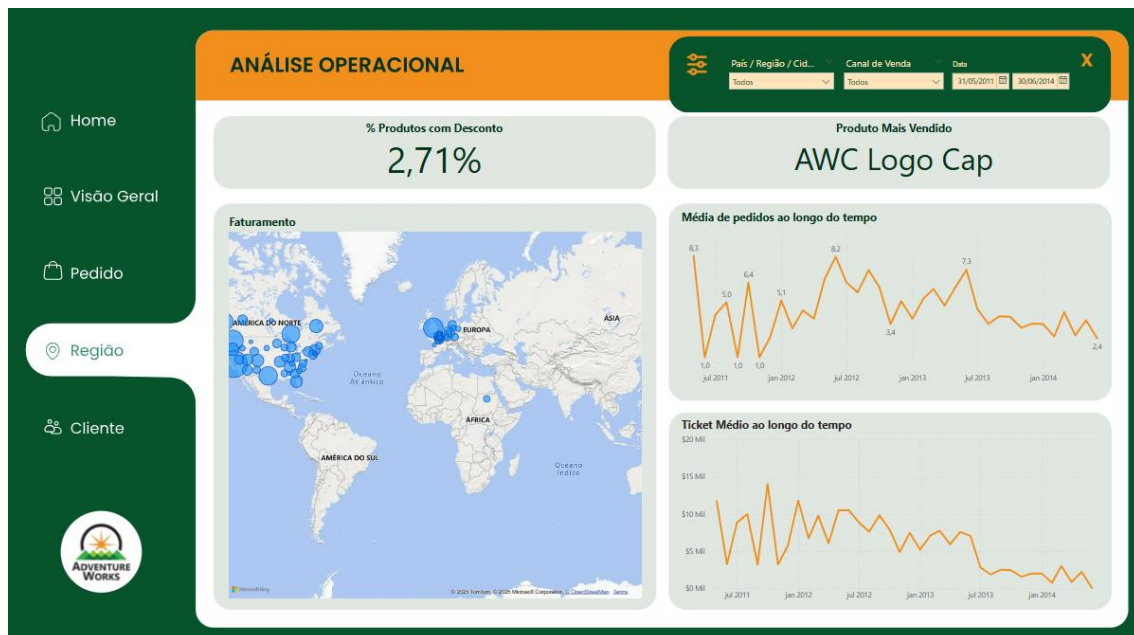
Aqui estão algumas perguntas de negócio que podem ser respondidas através das visualizações:

1. Qual é o impacto de produtos específicos (ex.: Bikes) no faturamento geral?
2. Como o ticket médio e o volume de vendas variam ao longo do tempo e entre categorias?
3. Quais métodos de pagamento são preferidos pelos clientes em diferentes regiões ou canais de venda?
4. Quais fatores motivam a maior parte das compras, e como isso influencia campanhas promocionais?
5. O Lifetime Value varia significativamente entre os clientes de diferentes canais de venda ou categorias de produtos?
6. Há uma correlação entre o número médio de produtos por pedido e o ticket médio ao longo do tempo?

Aqui se encontram alguns insights que foram possíveis de identificar:

- **Foco em "bikes":** Essa categoria é responsável por uma grande fatia do faturamento (\$95M). Investir em melhorias na cadeia de suprimentos ou promoções exclusivas pode gerar ainda mais receita.
- **Ticket médio em declínio:** Embora o volume de produtos vendidos esteja crescendo, o ticket médio está em queda. Pode ser necessário reavaliar estratégias de preços ou expandir produtos premium.
- **Preferências de pagamento:** A predominância de certos métodos de pagamento sugere que a oferta de condições facilitadas ou descontos específicos nesses métodos pode melhorar a conversão.
- **Preço como motivador principal:** O preço está claramente direcionando as decisões de compra. Estratégias como descontos agressivos ou promoções sazonais podem ser mais impactantes que campanhas focadas apenas em qualidade.
- **Produtos por pedido:** A média de 8,74 produtos por pedido indica uma propensão à compra em maior volume. Isso pode ser aproveitado com ofertas "compre mais, economize mais" ou kits promocionais.

8.4.3. Região



Aba de região

Esta aba foca em análises de vendas com base em localização geográfica, permitindo identificar tendências regionais, comportamento de compra ao longo do tempo e insights sobre o desempenho de produtos específicos.

→ Métricas

% de produtos com desconto (2,71%):

Indica a proporção de vendas realizadas com desconto. Útil para monitorar o impacto de promoções em diferentes regiões.

Produto mais vendido (AWC Logo Cap):

Destaque do produto com maior número de vendas, permitindo avaliar sua popularidade global.

→ Gráficos

Faturamento no mapa mundial:

Apresenta a distribuição de vendas por região:

América do Norte e Europa são os principais mercados, indicando um foco nessas localidades. Permite identificar mercados subexplorados para futuras campanhas.

Média de pedidos ao longo do tempo:

Mostra a variação na quantidade média de pedidos por mês: Picos em **julho de 2011** (8,3) e **julho de 2012** (8,2), com tendência de queda a partir de 2013. Pode estar relacionado a sazonalidade ou mudanças no comportamento de compra.

Ticket médio ao longo do tempo:

O ticket médio apresenta alta volatilidade, com quedas marcantes após **2012**. Indica a necessidade de entender quais fatores contribuíram para a queda no valor médio por compra.

Aqui estão algumas perguntas de negócio que podem ser respondidas através das visualizações:

1. Quais regiões contribuem mais para o faturamento global e quais têm potencial de crescimento?
2. A oferta de descontos está concentrada em regiões específicas?
3. Por que o "AWC Logo Cap" é o produto mais vendido e como seu sucesso varia entre regiões?
4. Qual é a relação entre a média de pedidos e o ticket médio em diferentes períodos?
5. O ticket médio é maior em regiões específicas ou está relacionado a promoções e descontos?
6. Como identificar sazonalidades regionais que influenciam nas vendas?

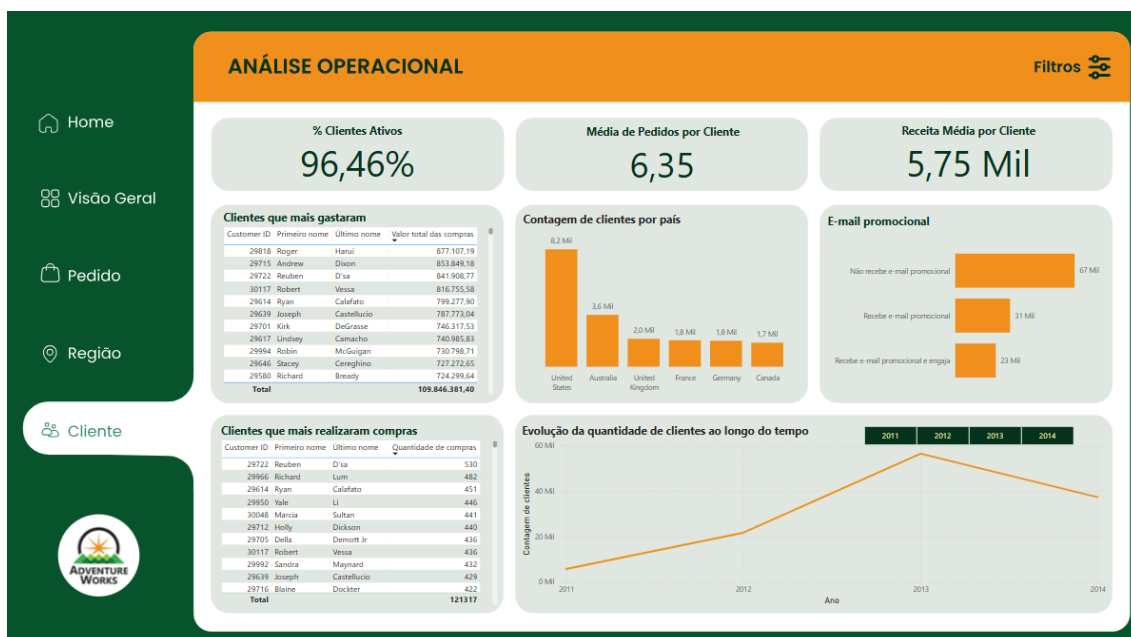
Após uma análise mais aprofundada, foi possível encontrar alguns insights:

- **Concentração de Faturamento em Mercados-Chave:** A América do Norte e a Europa são os mercados mais importantes. Estratégias específicas para fidelizar clientes nessas regiões devem ser priorizadas.
- **Produto Mais Vendido e Segmentação:** A popularidade do "AWC Logo Cap" pode ser utilizada em campanhas específicas, reforçando-o como produto principal para introduzir outros itens no mix de compras.
- **Queda no Ticket Médio:** A redução do ticket médio ao longo do tempo pode indicar aumento de produtos de baixo custo no

portfólio ou mudanças na percepção de valor. Reavaliar precificação e mix de produtos pode ser necessário.

- **Descontos Pouco Relevantes:** Apenas 2,71% dos produtos vendidos estão com desconto, sugerindo que promoções não são o principal motor das vendas. Isso pode ser ajustado caso a estratégia mude para atrair novos clientes.
- **Picos e Quedas na Média de Pedidos:** Picos em momentos específicos (como julho de 2011 e 2012) podem estar relacionados a campanhas sazonais. Estratégias semelhantes devem ser replicadas para estabilizar a queda recente.

8.4.4. Cliente



Aba de clientes

A aba de clientes possui informa  es e detalhamento sobre informa  es de clientes, condensando informa  es que pode auxiliar a compreender padr es no comportamento de compra.

→ M tricas

% Clientes ativos (96,46%):

Mede a propor  o de clientes que realizaram compras ou intera  es com a empresa em rela  o ao total de clientes cadastrados. Neste caso, ao modelar os dados, percebeu-se que h  clientes na tabela customer que n o fizeram compras ainda.

M dia de pedidos por cliente (6,35):

Indica o n mero m dio de compras realizadas por cliente ativo.

Receita M dia por Cliente (5,75 Mil):

Calcula o valor m dio gasto por cliente.

→ Gr ficos e Tabelas

Clientes que mais gastaram:

Exibe os clientes que geraram maior receita, com identificação por nome e valor total gasto.

Clientes que mais realizaram compras:

Lista os clientes com maior quantidade de pedidos realizados.

Contagem de Clientes por País:

Visualização da distribuição dos clientes em diferentes países, destacando os mercados mais representativos.

E-mail Promocional:

Representação de clientes que recebem e-mails promocionais, interagem ou não interagem com eles.

Evolução da Quantidade de Clientes ao Longo do Tempo:

Apresenta o crescimento ou declínio da base de clientes por ano.

Aqui estão algumas perguntas de negócio que podem ser respondidas através das visualizações:

1. Qual a eficiência da retenção e engajamento dos clientes ao longo do tempo?
2. Qual o comportamento de recompra dos clientes e como ele varia por região ou canal de venda?
3. Qual é o ticket médio por cliente e quais estratégias podem ser usadas para aumentá-lo?
4. Quem são os clientes de maior valor e como garantir a retenção desse grupo?
5. Quem são os clientes mais frequentes, e qual é o impacto do volume de compras na receita total?
6. Quais mercados apresentam maior concentração de clientes e quais devem ser priorizados em campanhas?
7. Qual a efetividade das campanhas por e-mail em engajar os clientes?
8. Quais períodos apresentaram maior crescimento na base de clientes, e quais fatores contribuíram para isso?

Após uma análise mais aprofundada da aba e a partir das perguntas de negócio, foi possível chegar aos seguintes insights:

- **Perfil de clientes valiosos:** Identificar padrões nos clientes que mais gastam ou realizam compras frequentes (ex.: localização, canal de vendas, histórico de interação).
- **Estratégias de engajamento:** Analisar a efetividade de campanhas por e-mail, otimizando o engajamento de clientes que não interagem com as comunicações.
- **Crescimento Sazonal:** O gráfico de evolução dos clientes pode revelar sazonalidade ou impactos de estratégias de marketing em determinados anos.

Com o dashboard, foi possível obter uma análise profunda de todos os setores que envolvem vendas e extrair insights. A interatividade entre os dados permite observar possíveis padrões e realizar planejamentos para as próximas campanhas, assim como entender a necessidade do mercado e direcionar os esforços para a obtenção dos resultados esperados.

9. Advanced Analytics

Para fazer o desenvolvimento da análise de previsão de demanda, foi escolhido o Google Colab, que é gratuito e permite o processamento de grandes volumes de dados.

9.1. Carregamento dos Dados

Os dados utilizados neste projeto foram obtidos através de uma query no BigQuery, que foi projetada para buscar apenas as informações relevantes para responder às perguntas propostas. Após executar a query, o resultado foi exportado em formato CSV para utilização no Colab.

```
with
customer as (
  select *
  from `adventureworks-desafio.adventureworks_desafio.stg_sap_adw__customer`
),
store as (
  select *
  from `adventureworks-desafio.adventureworks_desafio.stg_sap_adw__store`
),
orderheader as (
  select *
  from `adventureworks-desafio.adventureworks_desafio.stg_sap_adw__salesorderheader`
),
orderdetail as (
  select *
  from `adventureworks-desafio.adventureworks_desafio.stg_sap_adw__salesorderdetail`
),
address as (
  select *
  from `adventureworks-desafio.adventureworks_desafio.stg_sap_adw__address`
),
stateprovince as (
  select *
  from `adventureworks-desafio.adventureworks_desafio.stg_sap_adw__stateprovince`
),
countryregion as (
  select *
  from `adventureworks-desafio.adventureworks_desafio.stg_sap_adw__countryregion`
),
product as (
  select *
  from `adventureworks-desafio.adventureworks_desafio.dim_products`
),
salesterritory as (
  select
    stateprovince.stateprovincecode
    , stateprovince.stateprovince_name as state_name
    , stateprovince.stateprovinceid
    , countryregion.countryregioncode
    , countryregion.country_region_name as country_name
  from stateprovince
  left join countryregion ON stateprovince.countryregioncode = countryregion.countryregioncode
),
agg_prod_store AS (
  select
    customer.customerid
    , store.businessentityid
    , store_name
    , orderheader.salesorderid
    , orderheader.onlineorderflag
    , orderdetail.salesorderid
    , orderdetail.orderqty
    , orderdetail.unitprice
    , product.category_name
    , date(orderheader.orderdate) as orderdate
    , orderdetail.unitpricediscount
    , product.productid
    , product.product_name
    , salesterritory.stateprovincecode
    , salesterritory.state_name
    , salesterritory.country_name
    , orderheader.totaldue
  from customer
  left join store on customer.storeid = store.businessentityid
  left join orderheader on customer.customerid = orderheader.customerid
  left join orderdetail on orderdetail.salesorderid = orderheader.salesorderid
  left join product on orderdetail.productid = product.productid
  left join address on orderheader.billtoaddressid = address.addressid
  left join salesterritory on address.stateprovinceid = salesterritory.stateprovinceid
  where product.productid is not null
)
select * from agg_prod_store;
```

9.2. Carregamento, processamento e transformação dos dados

Já no colab, a primeira etapa feita foi o carregamento dos dados e a importação de bibliotecas que seriam utilizadas no decorrer do projeto.

Após essa etapa, foi feito o tratamento de dados faltantes. Durante a análise inicial, identificou-se a presença de dados faltantes na coluna `store_name`. Após analisar a coluna `onlineorderflag`, que indica compras realizadas online quando seu valor é `TRUE`, concluiu-se que os valores nulos em `store_name` correspondem a compras online. Esses valores foram substituídos por "e-commerce".

Uma outra transformação feita foi a alteração do tipo de dado na coluna `orderdate`, que foi convertido para o tipo `datetime`. Uma nova coluna foi criada para armazenar apenas o mês e o ano, visando facilitar as análises temporais futuras.

9.3. Modelagem e Redefinição de Índices

Para a etapa de modelagem, o índice do DataFrame foi redefinido com o valor 'total', transformado em uma nova coluna chamada `unique_id`, e o índice foi reiniciado de forma numérica.

9.4. Questões e Respostas

9.4.1. Questão 8

A fim de ajustar a distribuição de produtos e ter uma melhor estimativa sobre a necessidade de compra de matéria-prima, faça uma previsão sobre a demanda dos próximos três meses de cada produto em cada loja. Além disso, aponte se há ou não a presença de sazonalidade em algum produto de sua escolha.

A previsão foi realizada utilizando dois modelos:

1. Modelo SARIMA (Statsmodels):

- Configuração sazonal: (1, 1, 1, 12), indicando padrões mensais.
- Resultados da previsão (coluna `orderqty`):
 - Mês 35: 11.279,20
 - Mês 36: 17.017,57
 - Mês 37: 23.403,20

O modelo SARIMA identificou padrões de sazonalidade relacionados à demanda de produtos.

2. Modelo AutoARIMA (StatsForecast):

- Métricas de desempenho:
 - MAE: 8811,30
 - MSE: 142.434.496,00
 - RMSE: 11.934,59
 - MAPE: 348,50%

A alta taxa de erro percentual (MAPE) indica que o modelo AutoARIMA teve baixa acurácia e não é adequado para previsões de demanda.

9.4.2. Questão 9

Seria possível resolver este problema através de uma abordagem utilizando modelos de regressão? Se sim, qual demonstra melhor resultado? Justifique utilizando métricas de avaliação.

O modelo de Regressão Linear foi utilizado para prever a demanda usando mlforecast e LinearRegression

- Métricas de desempenho:
 - MAE: 9488,62
 - MSE: 156.883.664,00
 - RMSE: 12.525,32
 - MAPE: 366,56%

Esses resultados demonstram que o modelo de Regressão Linear não performou bem, com altos valores de erro absoluto e percentual. Isso sugere que esses modelos não foram adequados para capturar padrões não lineares ou sazonais presentes nos dados.

9.4.3. Questão 10

Os novos centros de distribuição passaram a ser divididos em províncias nos EUA e em países no resto do mundo. Qual desses grupos apresentou mais crescimento em demanda nos três meses previstos no item 8?

A análise dos gráficos mostra que o crescimento da demanda foi maior no grupo "resto do mundo" em comparação com as províncias dos EUA.

9.4.4. Questão 11

Um novo fornecedor de luvas, que agora engloba toda a produção mundial, precisa ter uma estimativa de quantos zíperes precisa pedir para os próximos 3 meses. Levando em consideração que são necessários 2 zíperes por par, quantos seriam necessários?

Com base na previsão de vendas, estima-se a produção de 643 pares de luvas por mês. Considerando que cada par requer 2 zíperes, a demanda total para os próximos três meses é de:

O fornecedor deve solicitar 3.858 zíperes para atender à produção global de luvas prevista para os próximos três meses.

10. Conclusão

Este projeto de construção de um pipeline de dados foi marcado pela sua complexidade e abrangência, abordando todas as etapas fundamentais do ciclo de dados. Ele proporcionou uma oportunidade valiosa para consolidar os conhecimentos adquiridos durante o programa Lighthouse, além de possibilitar a aplicação prática e o aprofundamento em temas que ainda não haviam sido explorados em profundidade.

Desafios dessa natureza são fundamentais para o desenvolvimento do profissional de dados, pois permitem vivenciar a integração de diversas habilidades técnicas e analíticas em um contexto prático. Ao longo do projeto, foi possível compreender melhor as demandas e os desafios envolvidos em processos reais, além de reforçar a importância de uma abordagem estruturada para resolver problemas complexos.