

Bases de Dados

Módulo 19: Baco de dados Analítico

Prof. André Bruno de Oliveira

28/05/24 10:03

Introdução DATA WAREHOUSE

- O que é um Data Warehouse (DW) ?
 - Uma DW pode ser definida como uma coleção de dados orientada a assunto, integrada, não volátil, variável no tempo para o suporte às decisões da gerência.
 - Os DW oferecem acesso a dados para análise complexa, descoberta de conhecimento e tomada de decisão. Eles dão suporte a demandas de alto desempenho sobre os dados e informações de uma organização.

1 Introdução DATA WAREHOUSE

- Entre as aplicações que interagem com o DW pode-se citar: OLAP e DSS e mineração de dados.
- OLAP é a abreviação de processamento analítico on-line, um termo usado para descrever a análise de dados DW. O OLAP exige utilização de computação distribuída, capacidade alta de armazenamento e um bom poder de processamento.
- DSS (Decision Support System - sistemas de apoio à decisão) São sistemas que fazem uso de dados de alto nível para apoiar decisões da liderança nas empresas.
- EIS (Executive Information Systems - Sistemas de Informação Executivos) Permite gerar relatórios com informações gerenciais a partir de aplicações de alto nível. É considerada uma forma especializada de DSS voltada para informações de gerenciais.
- A mineração de dados por exemplo é usada para busca de novo conhecimento não detectado nos dados.

1 Introdução DATA WAREHOUSE

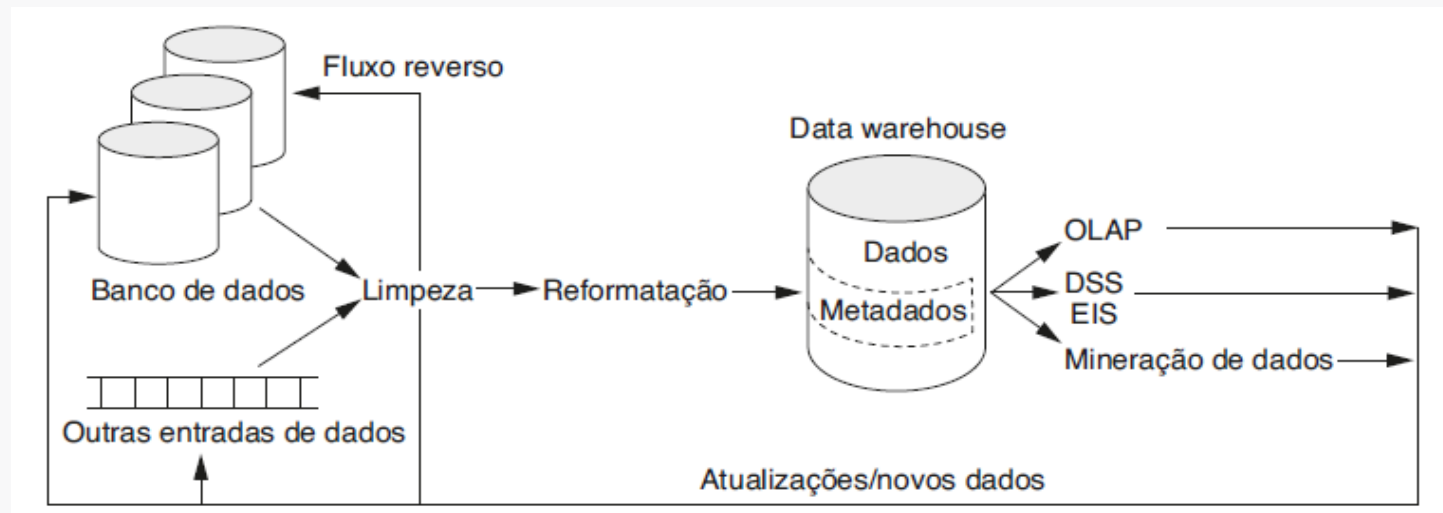
- Grosso modo os dados estão divididos em duas abordagens: i) Dados estruturados (relatórios e planilhas) e; ii) e informações sem clareza que precisam de técnicas matemáticas para descoberta da informação.
- Ex.:
 - Uso de modelos estatísticos para estabelecer relações entre os dados: Use de faixa etária combinado com a renda de uma população para determinar o grupo alvo sobre venda de cursos.
 - Uso de modelos preditivos sobre dados históricos para prever períodos de melhorar investimento. Por exemplo, venda de passagens aéreas.

1.2 Características dos data warehouses

- O DW faz uso do modelo de dados multidimensional que mantem um depósito de dados integrados de múltiplas fontes.
- Em comparação com os bancos de dados transacionais, os DW são não voláteis. Isso significa que as informações no DW mudam com muito menos frequência.
- Diferentemente dos bancos transacionais, DW trabalham com dados históricos. As informações têm um nível de detalhamento menor do que nos BD transacionais e atualizações são incrementais vinculadas a uma política da organização.
- Exemplo: Novas informações do mesmo produto são inseridas gerando duplicidades e uso de *surrogate* em tabelas para favorecer o poder processamento e a descoberta de informações relevantes.

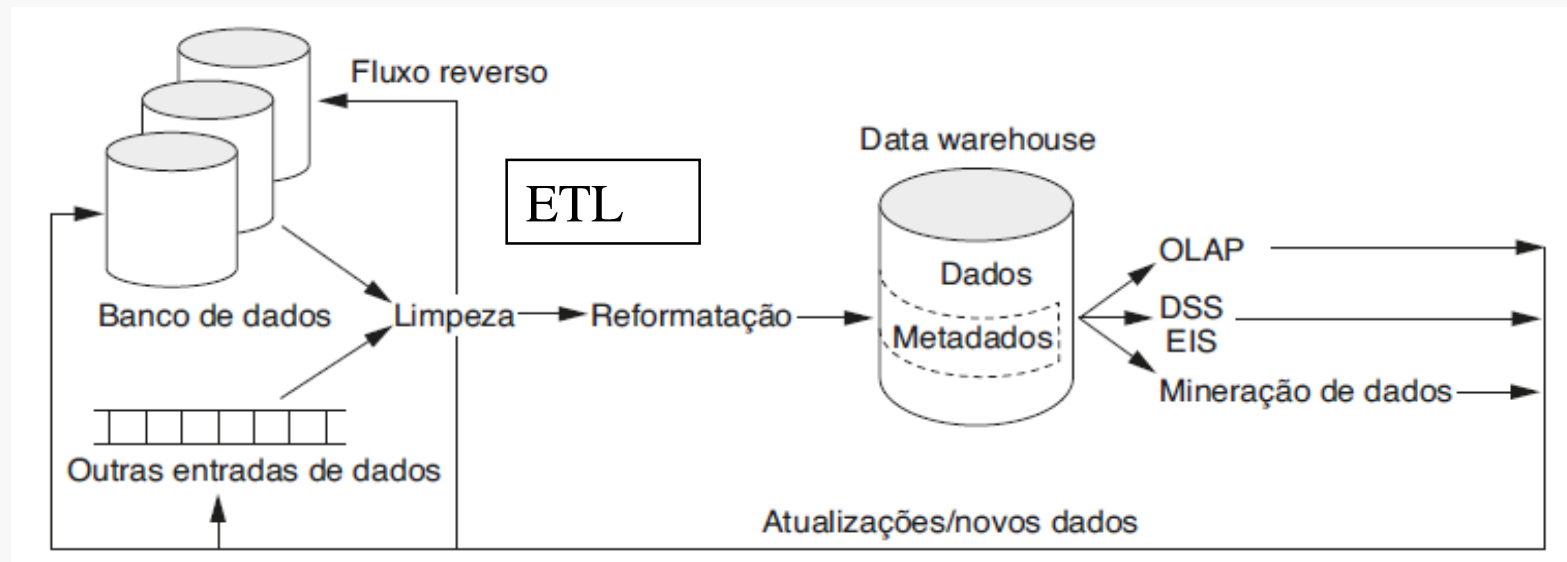
1.2 Características dos data warehouses

- A abaixo a figura oferece uma visão geral da estrutura conceitual de um DW.
- O processo inteiro inclui limpeza e reformatação dos dados antes que sejam carregados no DW. Este processo é tratado por soluções conhecidas como ferramentas de ETL (extração, transformação e carga).
- Perceba que no processo de elaboração de uma DW é fundamental reunir *softwares* capazes de transformar a informação para um esquema de linhas e colunas. Isso envolve o trabalho de profissionais capacitados.



1.2 Características dos data warehouses

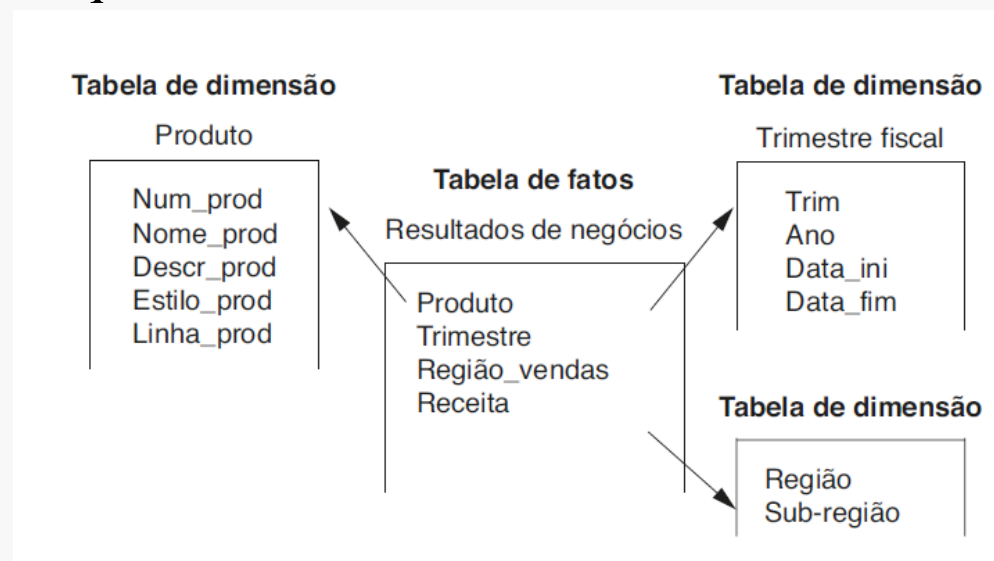
- No backend ocorre o processo de descoberta do conhecimento que podem gerar novas informações relevantes, como as regras de agregação envolvendo somatório, média, mínimo e máximo. Essas informações aparecem na figura num processo de retroalimentação (atualização de novos dados)
- O volume de dados chega na ordem de petabytes (10^{12}).



1.3 Modelagem de dados para DW

- Dois esquemas multidimensionais comuns são o **esquema estrela** e o **esquema floco de neve**. O esquema estrela consiste em uma tabela de fatos com uma única tabela para cada dimensão. A tabela fatos possui as FK de cada tabela de dimensão, numa relação de 1,N. Assim, uma tabela fatos possui muito mais registros do que a tabela de dimensão.
- Cada linha da tabela fatos contém valores resumos resultado de cruzamentos que consideram as características de cada linha da tabela dimensão, isso é feito na etapa de ETL. A inclusão de linhas nestas tabelas fatos e dimensão demanda estudo e planejamento.

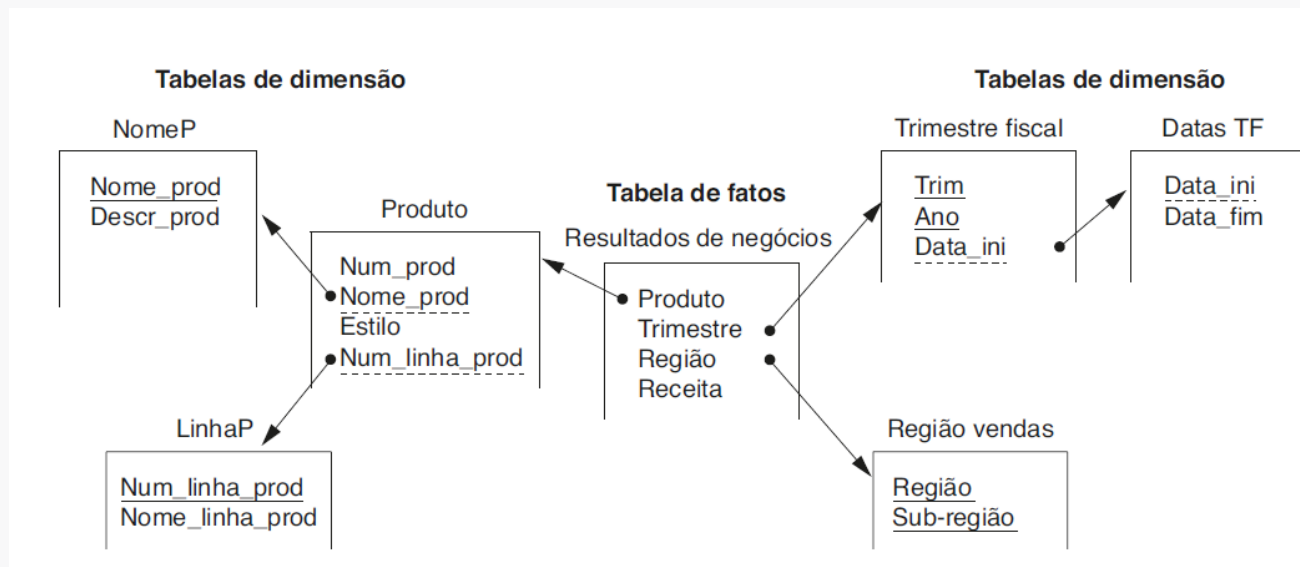
Esquema estrela



1.3 Modelagem de dados para DW

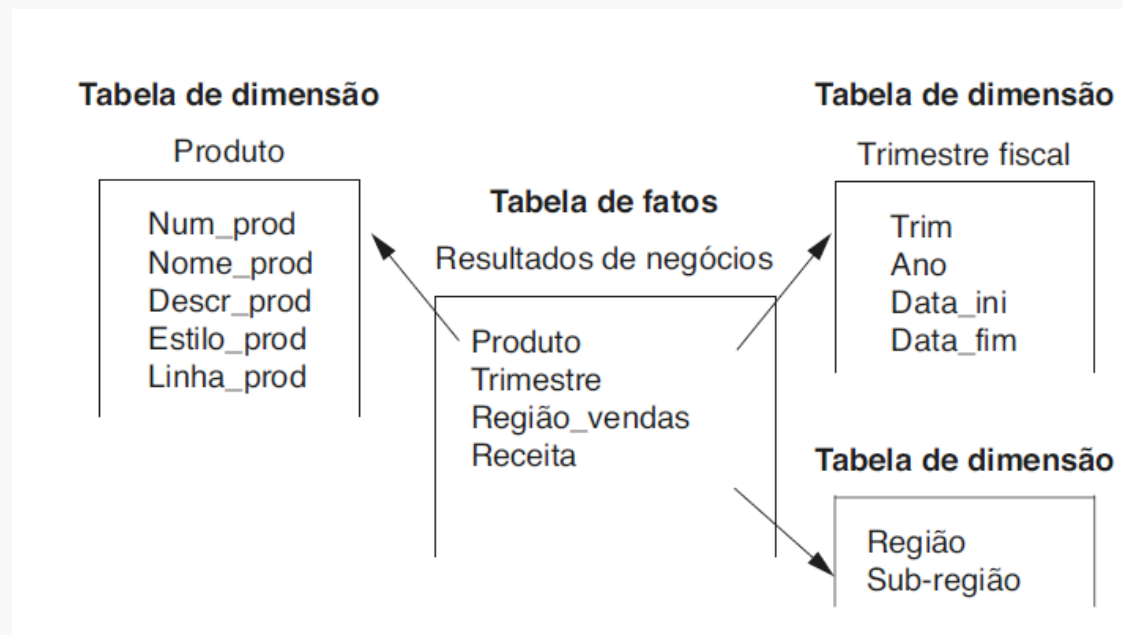
- O esquema floco de neve é uma variação do esquema estrela em que as tabelas de dimensões de um esquema estrela são organizadas em uma hierarquia para normalizá-las. A normalização visa reduzir as repetições que acabam aumentando o tamanho da tabela dimensão e impactando no espaço ocupado em HD, contudo o aumento de relacionamentos aumenta o custo das *queries* devido aos JOINS.

Esquema floco de neve



1.3 Modelagem de dados para DW

- O modelo de armazenamento multidimensional envolve dois tipos de tabelas: tabela dimensão e tabela fato.
- Uma tabela de dimensão consiste em tuplas de atributos da dimensão: Costumam usar *surrogate* para manter dados históricos.
- Uma tabela fato pode ser imaginada como tendo tuplas, uma para cada fato registrado. Cada tupla é composto pelo valor agregado e pelas FK de cada tabela dimensão.



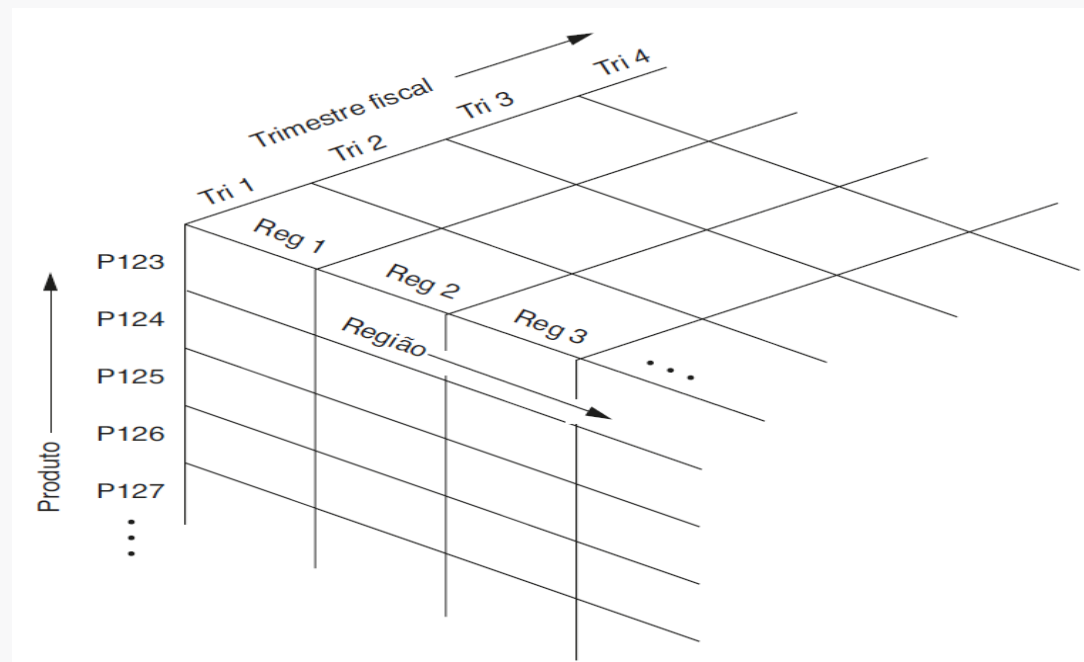
1.3 Modelagem de dados para DW

- O DW faz uso dos relacionamentos inerentes aos dados para gerar matrizes multidimensionais chamadas de cubos de dados quando se tem 3 dimensões. Quando há mais de 3 dimensões denomina-se hipercubos.
- Exemplo: Imagine uma planilha de vendas regionais por produto e região. Os produtos dispostos como linhas e as receitas de vendas para cada região compreendendo as colunas. O acréscimo de uma dimensão de tempo, como os trimestres fiscais de uma organização, produz uma matriz tridimensional, neste caso é possível representar estas dimensões usando um cubo de dados.

	Região			
	Reg 1	Reg 2	Reg 3	...
Produto	P123			
	P124			
	P125			
	P126			
	⋮			

1.3 Modelagem de dados para DW

- Veja a figura abaixo de cubo tridimensional: Um modelo com as dimensões, produto, receita de vendas por região e tempo por trimestre fiscal. Cada célula contém um produto e região específicas. Este tipo de modelo usa uma tecnologia que permite realizar consultas diretamente em qualquer combinação de dimensões, evitando consultas complexas diretamente no banco de dados.



1.4 Funcionalidade típica de um DW

- Os data DW existem para facilitar as consultas complexas, com uso intenso de dados. Neste sentido, precisam oferecer suporte à consulta muito maior e mais eficiente do que é exigido pelos bancos de dados transacionais.
- Entre as ferramentas e técnicas usadas estão no ROLAP, MOLAP:
- ROLAP (Relational On Line Analytical Processing) - Os dados são armazenados no BD relacional em estruturas multidimensionais e o SQL é utilizado para realizar as consultas.
- O MOLAP (Multidimensional On Line Analytical Processing) usa um cubo multidimensional que acessa os dados armazenados por meio de várias combinações. Os dados são pré-calculados, pré-resumidos e armazenados neste cubo, o que garante uma boa eficiência na análise dos dados.

1.4 Funcionalidade típica de um DW

- MOLAP oferece boa performance nas opções de *slicing and dicing* (quebrar um corpo de informações em partes menores ou examiná-lo de diferentes pontos de vista para que você possa compreendê-lo melhor).
- O ROLAP em comparação com o MOLAP é menos eficiente no tempo de resposta, porque ele armazena os dados em linhas e colunas das tabelas do BD e realiza junções, com isso o desempenho não é tão eficiente.

1.4 Funcionalidade típica de um DW

- O MOLAP é limitado na quantidade de dados que consegue guardar/manipular em memória. Além disso, adicionar dimensões nesta solução é mais complicado, ao invés disso costuma-se reconstruir as dimensões de dados.
- Pode-se dizer que os sistemas ROLAP são adequados para guardar grandes quantidades de dados, usando um processamento paralelo e tecnologias separadas, enquanto os sistemas MOLAP são adequados para aplicações departamentais com pequenos volumes de dados.

Bibliografia

Fontes de consulta:

<https://run.unl.pt/bitstream/10362/8403/1/TEGI0304.pdf>

<https://azure.microsoft.com/pt-br/resources/cloud-computing-dictionary/what-is-a-data-lake>

<https://azure.microsoft.com/pt-br/resources/cloud-computing-dictionary/what-is-a-data-warehouse>

<https://www.oracle.com/br/internet-of-things/what-is-iot/>

<https://www.dataside.com.br/dataside-community/big-data/data-lake-e-data-warehouse-do-conceito-a-arquitetura>

<https://azure.microsoft.com/pt-br/resources/cloud-computing-dictionary/what-is-a-data-lake>

Livros

Fundamentals of Database Systems, 7th Edition

Ramez Elmasri & Shamkant B. Navathe

Pearson, 2016.

O Livro Completo da Engenharia de Dados

Aprenda com casos de uso do mundo real

<https://www.databricks.com/br/resources/ebook/the-big-book-of-data-engineering> (Download gratuito).

Dissertação de mestrado em estatística. Soraia Vanessa Moura Velho. -

<https://run.unl.pt/bitstream/10362/8403/1/TEGI0304.pdf>

Prática sobre DW

- Exercício prático para a construção de um Data Warehouse (DW), focado no cenário de uma empresa de e-commerce. Esse exercício envolve a definição dos requisitos, modelagem, criação de tabelas e a aplicação de um processo simplificado de ETL.
- Cenário:
 - Você foi contratado por uma empresa de e-commerce para construir um data warehouse que ajudará a analisar as vendas de produtos para seus clientes, produtos e avaliar o desempenho de campanhas de marketing. A empresa utiliza várias fontes de dados, incluindo seu *sistema de vendas*, CRM (Customer Relationship Management – Gestão de Relacionamento com Cliente), e uma plataforma de marketing digital.

Prática sobre DW

- O objetivo é consolidar esses dados em um Data Warehouse e fornecer suporte para a criação de relatórios como:
 - Total de vendas por produto, cliente e região.
 - Performance de campanhas de marketing.
 - Tendências de compras ao longo do tempo.

Prática sobre DW

- Etapa 1: Definição dos Requisitos
- 1. Quais relatórios e métricas são importantes para a empresa?
 - Exemplo: Vendas por produto, região (cidade, estado), vendedor, análise de campanhas de marketing, análise de clientes recorrentes e novos.
- 2. Quais perguntas precisam ser respondidas pelo DW?
 - Quais produtos têm o maior número de *vendas por região*?
 - Qual foi o impacto de uma *campanha* específica nas vendas?
 - Quais são os *clientes* que mais compram?
 - Qual o *volume de vendas* por período (*mês, trimestre*)?

Prática sobre DW

- 3. Fontes de Dados:
 - ERP (dados de vendas - [quantidade, data] e produtos [nome do produto, preço, tipo ou categoria])
 - CRM (dados de clientes – nome, sexo, idade, cidade, estado)
 - Sistema de Marketing (dados das campanhas – nome da campanha, custo do investimento, canal de divulgação,)

ERP - Enterprise Resource Planning Significado:

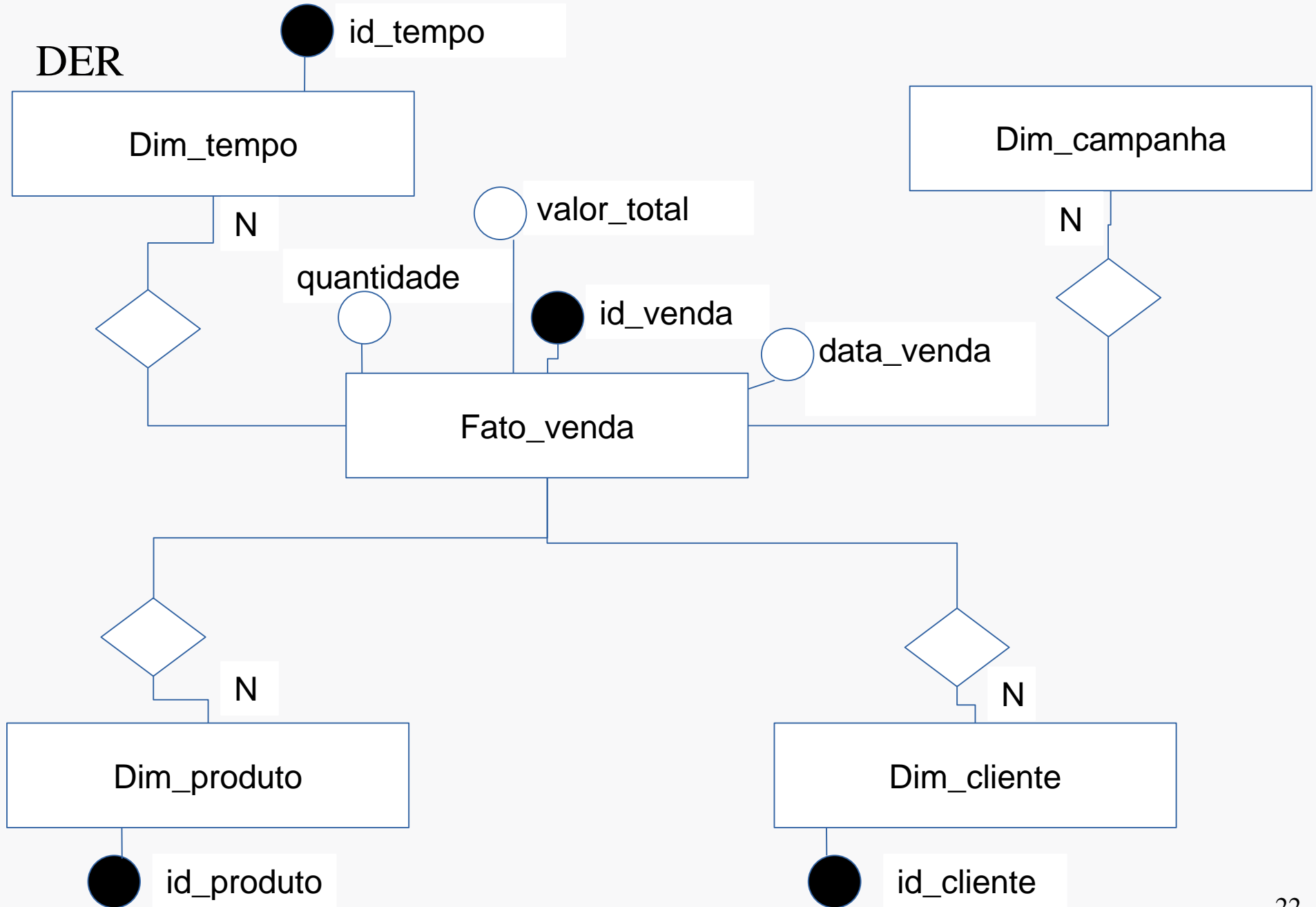
- Planejamento de Recursos Empresariais – é um sistema de informação que permite aos gestores de uma empresa terem, de forma integrada e simplificada, dados sobre os setores envolvidos nos processos de produção e/ou desenvolvimento de um produto ou serviço, indo desde o estoque e o fluxo de caixa até o pós-atendimento prestado aos clientes

Prática sobre DW

- **Etapa 2:** Modelagem do Data Warehouse
- Agora, você deve criar um modelo em estrela para organizar os dados no DW.
- Tabela Fato: Centraliza as métricas de interesse, neste caso, os dados de vendas.
- Tabelas Dimensão: Tabelas descritivas que fornecem detalhes sobre os atributos relacionados às vendas:
 - *Produto, Cliente, Tempo, Campanha*

Prática sobre DW

- DER



Prática sobre DW

- Detalhes

Dim_Produto

- ID_Produto
- Nome_Produto
- Categoria
- Preço

Dim_Cliente

- ID_Cliente
- Nome_Cliente
- Sexo (1- homem, 2- mulher)
- Idade inteiro
- Cidade (nome da cidade)

Dim_Campanha

- ID_Campanha
- Nome_Campanha
- Canal (Email, Redes Sociais, etc.)
- Custo_Campanha

Dim_Tempo

- ID_Tempo
- Dia
- Mês
- Ano
- Tempo

Prática sobre DW

- Projeto lógico

Dim_Produto:

- ID_Produto int (PK)
- Nome_Produto text
- Categoria text
- Preço real

Dim_Cliente

- ID_Cliente int (PK)
- Nome_Cliente text
- Sexo text check(sexo in ('M','F'))
- Idade int
- Cidade text

Dim_Campanha

- ID_Campanha int (PK)
- Nome_Campanha text
- Canal text
- Custo_Campanha real

Dim_Tempo

- ID_Tempo int (PK)
- Dia int
- Mês int
- Ano int
- Trimestre int

Prática sobre DW

- Projeto lógico

Fato_venda

- ID_venda int (PK)
- ID_Produto int (FK para dim_produto)
- ID_Cliente (FK para dim_cliente)
- ID_Campanha (FK para dim_campanha)
- ID_Tempo (FK para dim_tempo)
- Quantidade int
- valor_total real
- data_venda text

Prática sobre DW

- Projeto Físico

1. Criação da Tabela Dimensão: Dim_Produto

```
CREATE TABLE Dim_Produto (  
    ID_Produto INT PRIMARY KEY,  
    Nome_Produto text,  
    Categoria text,  
    Preco real);
```

2. Criação da Tabela Dimensão: Dim_Cliente

```
CREATE TABLE Dim_Cliente (  
    ID_Cliente INT PRIMARY KEY,  
    Nome_Cliente text,  
    Sexo text check( sexo in ('M','F')),  
    Idade int,  
    Cidade text,  
    Estado text);
```

Prática sobre DW

- Projeto Físico

3. Criação da Tabela Dimensão: Dim_Campanha

```
CREATE TABLE Dim_Campanha (  
    ID_Campanha INT PRIMARY KEY,  
    Nome_Campanha text,  
    Canal text,  
    Custo_Campanha real);
```

4. Criação da Tabela Dimensão: Dim_Tempo

```
CREATE TABLE Dim_Tempo (  
    ID_Tempo INT PRIMARY KEY,  
    Dia INT,  
    Mes INT,  
    Ano INT,  
    Trimestre INT);
```

Prática sobre DW

- Projeto Físico

5. Criação da Tabela Fato: Fato_Vendas

```
CREATE TABLE Fato_Vendas (  
    ID_Venda INT PRIMARY KEY,  
    ID_Produto INT,  
    ID_Cliente INT,  
    ID_Campanha INT,  
    ID_Tempo INT,  
    Quantidade INT,  
    Valor_Total real,  
    Data_Venda text,  
    FOREIGN KEY (ID_Produto) REFERENCES Dim_Produto(ID_Produto),  
    FOREIGN KEY (ID_Cliente) REFERENCES Dim_Cliente(ID_Cliente),  
    FOREIGN KEY (ID_Campanha) REFERENCES Dim_Campanha(ID_Campanha),  
    FOREIGN KEY (ID_Tempo) REFERENCES Dim_Tempo(ID_Tempo)  
);
```

Prática sobre DW

- **Etapa 3: Processo de ETL (Extração, Transformação e Carga)**
- Nesta etapa, você implementará o processo de ETL para mover os dados das fontes de origem para o Data Warehouse.
- 1. Extração: Coletar os dados brutos do ERP, CRM e sistema de marketing.
 - Extraia os dados de vendas do ERP, que contêm informações sobre as transações (produto, quantidade, valor, data).
 - Extraia os dados dos clientes do CRM, contendo informações pessoais (nome, localização, etc.).
 - Extraia os dados das campanhas do sistema de marketing, contendo as campanhas lançadas e os canais de distribuição.

Prática sobre DW

- **Etapa 3:** Processo de ETL (Extração, **Transformação** e Carga)
- 2. Transformação: Padronize e limpe os dados.
 - Combine as informações de produtos e clientes, removendo duplicatas.
 - Ajuste o formato das datas e valores monetários.
 - Garanta que todas as chaves estrangeiras estejam corretas (ex.: cada venda tem um ID_Cliente válido).

Prática sobre DW

- **Etapa 3:** Processo de ETL (Extração, Transformação e **Carga**)
- 3. Carga: Carregue os dados transformados nas tabelas do Data Warehouse.
 - Carregue as vendas na Tabela Fato_Vendas.
 - Carregue os dados de produtos na Dim_Produto, clientes na Dim_Cliente, campanhas na Dim_Campanha e datas na Dim_Tempo.

Prática sobre DW

- **Etapa 3:** Processo de ETL (Extração, Transformação e **Carga**)

1. Inserção de Dados na Tabela Dimensão: Dim_Produto

```
INSERT INTO Dim_Produto (ID_Produto, Nome_Produto, Categoria,  
Preco)
```

```
VALUES
```

```
(101, 'Camiseta', 'Vestuário', 50.00),
```

```
(102, 'Celular', 'Eletrônicos', 150.00),
```

```
(103, 'Notebook', 'Eletrônicos', 150.00),
```

```
(104, 'Tênis', 'Calçados', 100.00),
```

```
(105, 'Fone de Ouvido', 'Eletrônicos', 25.00);
```


Prática sobre DW

- **Etapa 3:** Processo de ETL (Extração, Transformação e **Carga**)

2. Inserção de Dados na Tabela Dimensão: Dim_Cliente

```
INSERT INTO Dim_Cliente (ID_Cliente, Nome_Cliente, Sexo, Idade, Cidade, Estado)
```

```
VALUES
```

```
(201, 'João Silva', 'M', 34, 'São Paulo', 'SP'),
```

```
(202, 'Maria Oliveira', 'F', 28, 'Rio de Janeiro', 'RJ'),
```

```
(203, 'Ana Souza', 'F', 45, 'Belo Horizonte', 'MG'),
```

```
(204, 'Carlos Lima', 'M', 22, 'Curitiba', 'PR'),
```

```
(205, 'Luiza Martins', 'F', 39, 'Porto Alegre', 'RS');
```

Prática sobre DW

- **Etapa 3:** Processo de ETL (Extração, Transformação e **Carga**)

3. Inserção de Dados na Tabela Dimensão: Dim_Campanha

```
INSERT INTO Dim_Campanha (ID_Campanha, Nome_Campanha,  
Canal, Custo_Campanha)
```

```
VALUES
```

```
(301, 'Black Friday 2023', 'Email', 10000.00),
```

```
(302, 'Natal 2023', 'Redes Sociais', 8000.00),
```

```
(303, 'Aniversário E-commerce', 'SMS', 5000.00);
```

Prática sobre DW

- **Etapa 3:** Processo de ETL (Extração, Transformação e **Carga**)

4. Inserção de Dados na Tabela Dimensão: Dim_Tempo

```
INSERT INTO Dim_Tempo (ID_Tempo, Dia, Mes, Ano, Trimestre)
```

```
VALUES
```

```
(401, 15, 1, 2023, 1),
```

```
(402, 10, 2, 2023, 1),
```

```
(403, 20, 3, 2023, 1),
```

```
(404, 5, 4, 2023, 2),
```

```
(405, 25, 5, 2023, 2);
```

Prática sobre DW

- **Etapa 3:** Processo de ETL (Extração, Transformação e **Carga**)

5. Inserção de Mais Dados na Tabela Fato: Fato_Vendas

```
INSERT INTO Fato_Vendas (ID_Venda, ID_Produto, ID_Cliente,  
ID_Campanha, ID_Tempo, Quantidade, Valor_Total, Data_Venda)
```

```
VALUES
```

```
(6, 102, 201, 303, 401, 3, 450.00, '2023-06-12'),
```

```
(7, 105, 202, 301, 402, 1, 25.00, '2023-07-22'),
```

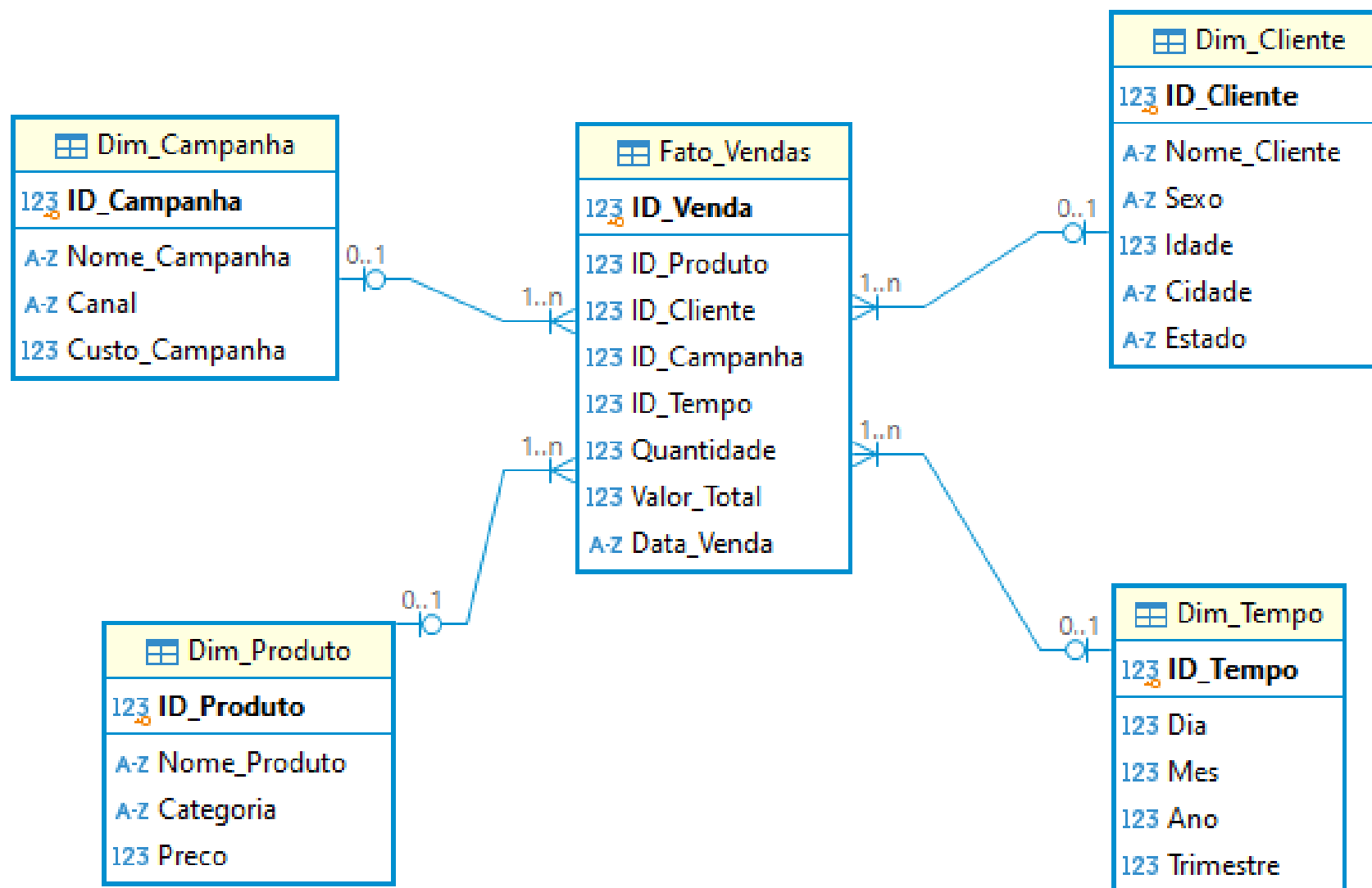
```
(8, 101, 203, NULL, 403, 2, 100.00, '2023-08-30'),
```

```
(9, 104, 204, 302, 404, 4, 400.00, '2023-09-15'),
```

```
(10, 103, 205, 303, 405, 1, 150.00, '2023-10-18')
```

DATA WHAREHOUSE ECOMMERCE – GERADO PELO DBEAVER

Veja que a cardinalidade está trocada.



Prática sobre DW

- **Etapa 4:** Criação de Relatórios e Dashboards (área de trabalho)
- Com os dados organizados no Data Warehouse, crie relatórios para responder às perguntas de negócios.
 - 1. Total de Vendas por Produto e Cidade, Estado:

Agrupe os dados da Fato_Vendas por produto e região (dimensões Dim_Produto e Dim_Cliente).

Create View Total_vendas_cidade_estado as

```
SELECT p.Nome_Produto, c.cidade, c.estado, SUM(f.Quantidade) AS Total_Vendas  
FROM Fato_Vendas f  
INNER JOIN Dim_Produto p ON f.ID_Produto = p.ID_Produto  
INNER JOIN Dim_Cliente c ON f.ID_cliente=C.id_cliente  
GROUP BY p.Nome_Produto, c.cidade, c.estado;
```

Prática sobre DW

- **Etapa 4:** Criação de Relatórios e Dashboards (área de trabalho)
- 2. Vendas por Cliente e Cidade:
 - Create view Cliente_cidade as
 - SELECT c.Nome_Cliente, c.Cidade, SUM(f.Valor_Total) AS Total_Gasto
 - FROM Fato_Vendas f
 - JOIN Dim_Cliente c ON f.ID_Cliente = c.ID_Cliente
 - GROUP BY c.Nome_Cliente, c.Cidade;

Prática sobre DW

- **Etapa 4:** Criação de Relatórios e Dashboards (área de trabalho)
- 3. Desempenho de Campanhas de Marketing:

Create view Marketing as

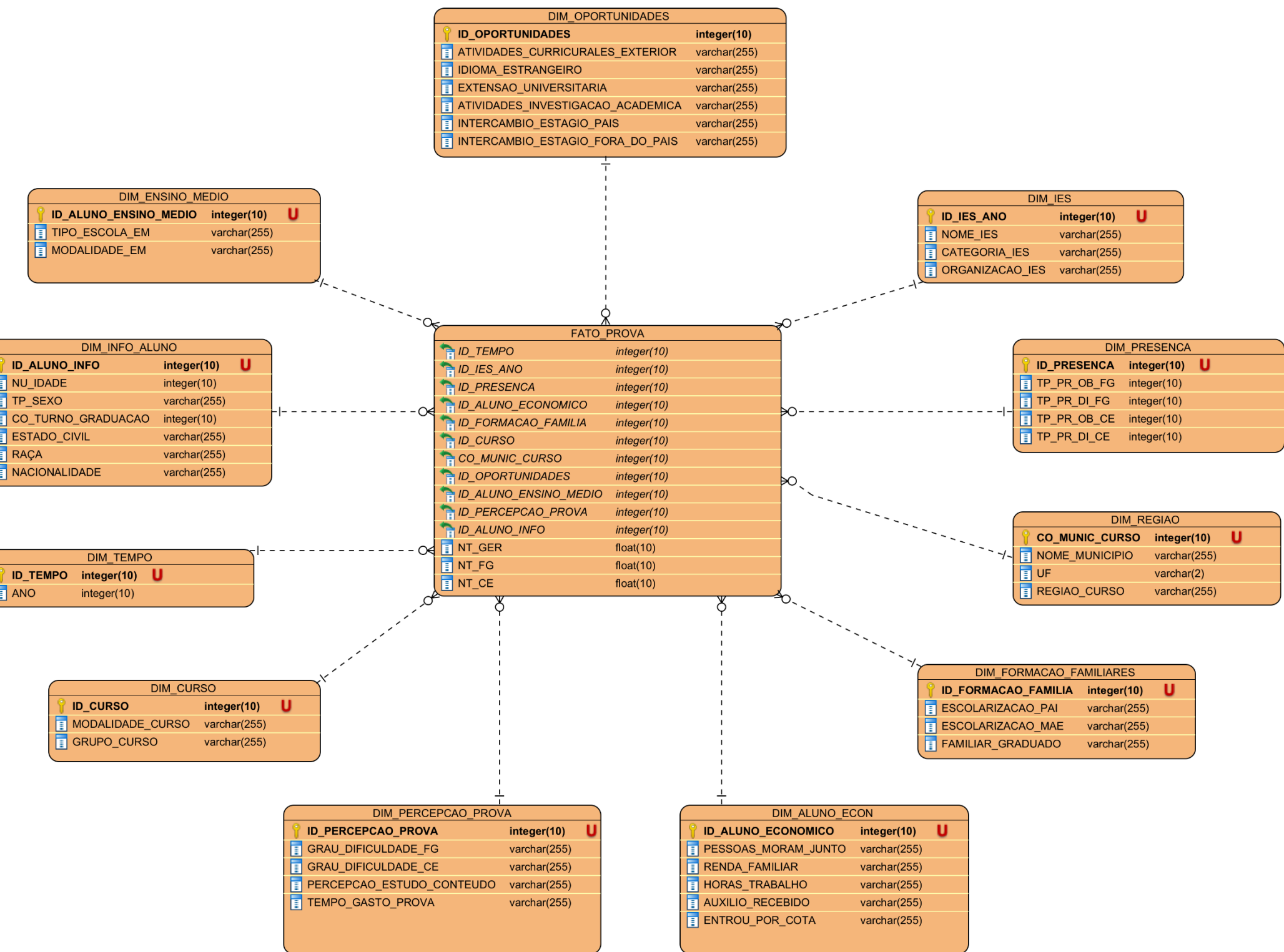
- `SELECT camp.Nome_Campanha, SUM(f.Valor_Total) AS Total_Vendas`
- `FROM Fato_Vendas f`
- `JOIN Dim_Campanha camp ON f.ID_Campanha = camp.ID_Campanha`
- `GROUP BY camp.Nome_Campanha;`

Prática sobre DW

- **Etapa 4:** Criação de Relatórios e Dashboards (área de trabalho)
- 4. Tendência de Vendas ao Longo do Tempo:
 - Create view Vendas_tempo as
 - SELECT t.Mes, t.Ano, SUM(f.Valor_Total) AS Vendas_Totais
 - FROM Fato_Vendas f
 - JOIN Dim_Tempo t ON f.ID_Tempo = t.ID_Tempo
 - GROUP BY t.Mes, t.Ano
 - ORDER BY t.Ano, t.Mes;

DW DO ENADE

- DW criado a partir dos microdados do Enade disponibilizados pelo site do INEP. DW está baseado em dados de todos os estudantes que realizaram o exame no ano de 2017, 2018 e 2019. Criado pela Letícia Tavares (Fonte: https://github.com/leticiatavaresds/Projeto_ENADE?tab=readme-ov-file).
- No presente trabalho foi utilizada a modelagem Estrela (Star Schema) caracterizada por poucas tabelas e relacionamentos, onde todas as tabelas de Dimensão se relacionam direta e unicamente com a tabela Fato, sendo assim um modelo simples e eficiente.
- O modelo elaborado neste trabalho foi desenvolvido com o intuito de ser centrado nas notas das provas (NT_FG – formação geral, NT_CE – conhecimento específico, NT_GER – nota geral), objetivando fornecer dados para análise de fatores que influenciam no desempenho de cada aluno que presta a avaliação. Dessa forma, o modelo apresenta como fato a prova que possui como métrica a nota final e as notas de formação geral incluindo seu componente específico, além de 11 tabelas de dimensão.



Exercícios práticos - DW DO ENADE

- (1) Encontre a média da nota geral (FATO) para cada grupo de cursos (DIM_CURSO) e exiba as médias na ordem decrescente .
- (2) Encontre os cinco primeiros grupos de cursos de maior média geral.
- (3) Encontre a média geral do grupo do curso de medicina por sexo e raça (DIM_INFO_ALUNO) e exiba as médias na ordem decrescente.
- (4) Retorne o total de candidatos em cada instituição de ensino (DIM_IES) para o ano de 2019. O resultado deve exibir a instituição com menor número de candidatos primeiro.
- (5) Retorna uma lista contendo instituições de ensino que não tiveram candidatos em 2019.

Obrigado