Data Warehousing

Arquiteturas, ETL e Estratégia de Dados

Camila Paranhos **Filipe Rubson** **Patrícia Silva** **Valdir Neto**

IFMG *Campus* São João Evangelista

08/12/2025

Agenda do Seminário

1. Definição e Conceitos Fundamentais

2. DW vs. Banco Operacional (OLAP vs OLTP)

3. Processo de Construção (ETL)

4. Modelagem Dimensional

5. Aplicações e Benefícios

6. Exemplos Práticos

7. Questões da Prova

Referências

O Paradoxo dos Dados

“Estamos nos afogando em dados, mas famintos por conhecimento.”

— John Naisbitt

O Problema: Sistemas operacionais geram terabytes de dados, mas eles estão fragmentados e focados na transação, não na análise.

A Solução: Data Warehouse (DW)

Um ambiente arquitetado para transformar dados brutos em inteligência estratégica.

Definição Canônica (Bill Inmon)

Um Data Warehouse é uma coleção de dados:

- **Orientada por Assunto**
- **Integrada**
- **Não Volátil**
- **Variável no Tempo**

Objetivo: Apoiar a tomada de decisões gerenciais.

1.1 Orientado por Assunto

Sistemas Operacionais:

- Organizados por *Processos* (Vendas, Estoque, Folha).
- Foco na função da aplicação.

Data Warehouse:

- Organizado por *Entidades* (Cliente, Produto, Região).
- Foco no objeto de análise.

Exemplo Intuitivo: A Biblioteca

Operacional: Livros organizados por Editora (quem produziu).
DW: Livros organizados por Assunto (História, Ficção).

1.2 Integrado (O Maior Desafio)

Sistemas diferentes ‘falam’ línguas diferentes. O DW é o tradutor.

Sistema A (RH)	**Sistema B (Vendas)**	**Data Warehouse**
Sexo: “M” / “F”	Sexo: 1 / 0	Sexo: “Masculino”
Data: 13/01/25	Data: 2025-01-13	Data: 2025-01-13
ID: Emp_001	ID: Vendedor_99	ID: Staff_Key_10

Objetivo: Single Version of Truth (SVOT).

1.3 Não Volátil (Histórico)

Conceito:

- Dados no DW são *Read-Only*.
- Não se altera o passado (***Update***), apenas se insere novas informações.
- Permite auditoria e análise de evolução.

Analogia: Espelho vs. Álbum

OLTP (Espelho): Mostra o “agora”. Se cortar o cabelo, a imagem muda.
DW (Álbum): Guarda a foto de 2010, 2015 e 2025. Preserva a história.

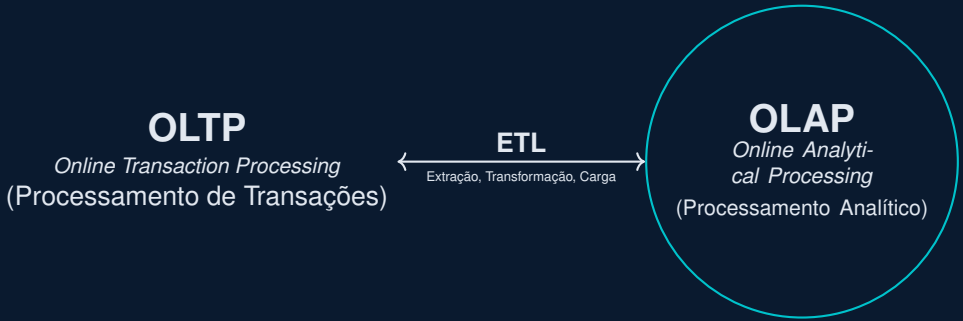
1.4 Variável no Tempo

Todo registro no DW possui um carimbo de tempo.

- Horizonte Operacional: 60 a 90 dias (para performance).
- Horizonte DW: 5 a 10 anos (para tendências).

Por que importa? Sem histórico longo, não há *Machine Learning* preditivo nem análise de sazonalidade.

A Grande Divisão: OLTP vs. OLAP



OLTP (Operacional)

Focado em **registrar** o dia a dia. É otimizado para operações rápidas e curtas (Ex: Passar um cartão de crédito).

OLAP (Decisório)

Focado em **analisar** os dados. É otimizado para ler milhões de registros para gerar um gráfico.

Comparativo Detalhado

Característica	OLTP (Operacional)	OLAP (DW)
Foco	Transação (Venda unitária)	Análise (Tendência global)
Volatilidade	Alta *(Write-Intensive)*	Baixa *(Read-Intensive)*
Estrutura	Normalizada (3NF)	Desnormalizada (Star Schema)
Performance	Milissegundos	Segundos/Minutos

Entendendo os termos técnicos:

- **Volatilidade (Write vs Read):** O OLTP sofre muitas escritas (inserções/atualizações) a todo segundo. O OLAP é estático, sofrendo alterações massivas apenas durante a Carga (ETL).
- **Normalizada (3NF):** Dados quebrados em muitas tabelas para evitar repetição (economiza espaço, mas exige muitos JOINs). Ideal para OLTP.
- **Desnormalizada (Star):** Dados repetidos propositalmente para facilitar a leitura. Otimiza a velocidade de consulta para relatórios. Ideal para OLAP.

Diferença nas Queries (SQL)

Query OLTP (Simples)

```sql
UPDATE Estoque
SET Qtd = Qtd - 1
WHERE ProdID = 555;
-- Toca em 1 linha.
-- Deve ser atomica.
```

Query OLAP (Pesada)

```sql
SELECT Regiao, SUM(Vendas)
FROM Fato_Vendas f
JOIN Dim_Tempo t ON...
WHERE t.Ano IN (2024, 2025)
GROUP BY Regiao;
-- Varre milhoes de linhas.
-- Agrega e Soma.
```

Por que separar os ambientes?

Risco de Performance

Se rodarmos a query OLAP (pesada) no banco OLTP (onde ocorrem as vendas), causamos **travamento (locks)** nas tabelas. O cliente não consegue comprar porque o analista está rodando um relatório.

O DW protege a operação diária das cargas de análise.

O Pipeline de Dados: ETL

Extract, Transform, Load: O motor do Data Warehouse.

Limpeza e Padronização

Etapa 1: Extract (Extração)

Desafio: Retirar dados sem derrubar o sistema de origem.

- **Carga Full:** Extrai tudo. (Usado na carga inicial).
- **Carga Incremental (Delta):** Extrai apenas o que mudou desde a última carga.
  - Usa Timestamps (*updated_at*).
  - Usa CDC (*Change Data Capture*) lendo logs do banco.

Etapa 2: Transform (Limpeza)

Onde a “sujeira” é tratada.

Problema	Dado Sujo	Dado Limpo
Formato	13/01/25	2025-01-13
Padronização	“S. Paulo”	“São Paulo”
Nulos	Venda: NULL	Venda: 0.00
Duplicidade	“J. Silva”	“João Silva” (ID 101)

Etapa 2: Transform (Enriquecimento)

Não é só limpar, é agregar valor.

- **Derivação:** Calcular “Lucro” (Venda - Custo).
- **Enriquecimento:** Pegar o CEP do cliente e adicionar dados de Renda Média do Bairro (fonte externa).
- **Codificação:** Converter códigos de produto legados para SKUs globais.

Etapa 3: Load (Carga e SCD)

Como gerenciar mudanças no histórico? (Slowly Changing Dimensions - SCD).

Cenário: Cliente Maria muda de Solteira para Casada.

- **SCD Tipo 1 (Sobrescrita):** Atualiza o registro. Perde-se o fato de que ela comprou como Solteira antes.
- **SCD Tipo 2 (Histórico - Padrão DW):** Cria uma nova linha.
  - Linha 1: Maria | Solteira | Data_Fim: Ontem
  - Linha 2: Maria | Casada | Data_Inicio: Hoje

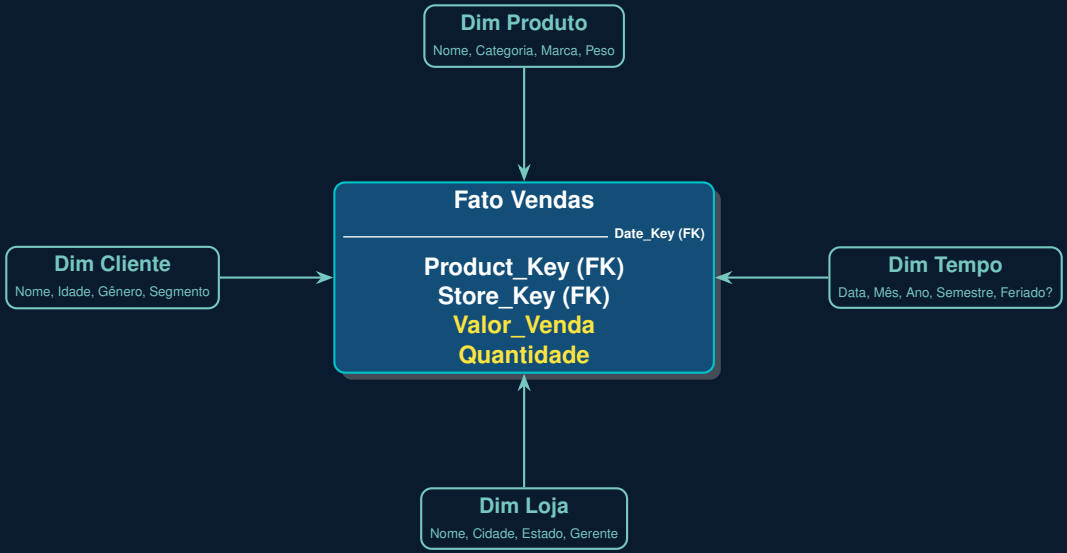
Modelagem Dimensional (Kimball)

Diferente da modelagem relacional (ER) focada em não repetir dados, o DW foca em **facilidade de consulta**.

Temos dois conceitos:

- **Fato (Fact):** O centro. Contém números (métricas).
- **Dimensão (Dimension):** As pontas. Contém texto (contexto).

O Esquema Estrela (Star Schema)



Entendendo Fato vs. Dimensão

Regra de Ouro

- **Fato = Verbos (Eventos):** Vendeu, Clicou, Ligou, Sacou. São números que queremos somar. Ninguém soma “CPFs” ou “Endereços”, então eles não são fatos.
- **Dimensão = Substantivos (Contexto):** Quem? Quando? Onde? O Quê? São filtros que usamos para agrupar no relatório.

Exemplo (Fato): O cliente ***comprou***, o aluno se ***matriculou***, o paciente ***foi internado***.

Exemplo (Dimensão): Quero ver as vendas (Fato) ***por*** Mês (Dimensão) e ***por*** Loja (Dimensão).

Star Schema vs. Snowflake Schema

Star Schema:

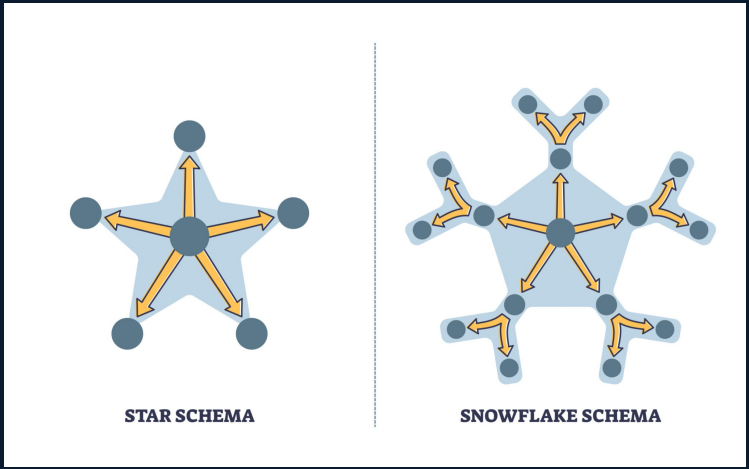
- Dimensões desnormalizadas (Cidade e Estado na mesma tabela).
- Mais rápido para ler (menos Joins).

Snowflake Schema:

- Dimensões normalizadas (Dim Cidade $\rightarrow$ Dim Estado).
- Economiza espaço, mas é mais lento para consultar.

No Data Warehouse, geralmente prefere-se o Star Schema. Embora o Snowflake economize espaço, no mundo da análise de dados (OLAP), prioriza-se a velocidade de leitura, e o Star Schema é mais simples e rápido para gerar relatórios.

Star Schema vs. Snowflake Schema (Visual)



Aplicações por Setor: Varejo

Market Basket Analysis (Cesta de Compras)

- *Pergunta:* “Clientes que compram fraldas também compram cerveja?”
- *Ação:* Colocar produtos próximos ou criar promoções conjuntas.
- *Benefício:* Aumento do ticket médio e otimização de estoque.

Aplicações por Setor: Finanças

Detecção de Fraude e Risco

- *Pergunta:* “Esta transação foge do padrão histórico deste cliente?”
- *Ação:* Bloqueio preventivo do cartão.
- *Conformidade:* Relatórios regulatórios (Basel III) exigem histórico consolidado e auditável que só o DW provê.

Aplicações por Setor: Saúde

Gestão Hospitalar e Clínica

- *Pergunta:* “Qual a taxa de readmissão de pacientes tratados com o protocolo X?”
- *Ação:* Ajuste de protocolos médicos baseados em evidências.
- *Benefício:* Melhora no cuidado ao paciente e redução de custos operacionais.

Cenário Prático: E-commerce Dark Data

Empresa: Loja de componentes de PC.

Problema: O sistema de vendas trava na Black Friday se rodarmos relatórios.

Solução DW:

- Extração incremental a cada 1 hora dos pedidos.
- Transformação: Calcular margem de lucro por peça.
- Carga: Star Schema no Google BigQuery.
- Resultado: Dashboard em tempo real sem afetar a venda.

Benefícios Estratégicos Gerais do DW

1. **Velocidade:** Consultas em minutos, não dias.
2. **Consistência:** “Single Source of Truth”. O Financeiro e o Vendas usam os mesmos números.
3. **Histórico:** Capacidade de prever o futuro analisando o passado (Forecasting).
4. **Self-Service BI:** Empoderar o usuário final para criar seus próprios gráficos.

Questão 1: Características do DW

1. Segundo Inmon, qual característica garante que os dados históricos não sejam sobrescritos, permitindo auditoria?

A Não Volátil
B Integrado
C Orientado por Assunto
D Variável no Tempo
E Normalizado

Questão 1: Características do DW

1. Segundo Inmon, qual característica garante que os dados históricos não sejam sobrescritos, permitindo auditoria?

A Não Volátil
B Integrado
C Orientado por Assunto
D Variável no Tempo
E Normalizado

Resposta Correta: A

A não volatilidade impede alterações (updates) em dados passados. O DW é fundamentalmente *read-only* após a carga.

Questão 2: OLTP vs OLAP

2. Qual a principal diferença de propósito entre OLTP e OLAP?

A OLAP é para inserção rápida; OLTP é para leitura.
B OLTP armazena histórico; OLAP armazena o dia atual.
C OLTP foca na transação operacional; OLAP foca na análise gerencial.
D Não há diferença, depende apenas do software usado.

Questão 2: OLTP vs OLAP

2. Qual a principal diferença de propósito entre OLTP e OLAP?

A OLAP é para inserção rápida; OLTP é para leitura.
B OLTP armazena histórico; OLAP armazena o dia atual.
C OLTP foca na transação operacional; OLAP foca na análise gerencial.
D Não há diferença, depende apenas do software usado.

Resposta Correta: C

OLTP sustenta a operação diária (vender). OLAP sustenta a decisão (analisar a venda).

Questão 3: Processo ETL

3. Na etapa de Transformação (T) do ETL, qual atividade NÃO é comum?

A Limpeza de duplicatas.
B Padronização de datas (ISO).
C Enriquecimento de dados (adicionar CEP).
D Criação de novos pedidos de venda no sistema de origem.

Questão 3: Processo ETL

3. Na etapa de Transformação (T) do ETL, qual atividade NÃO é comum?

A Limpeza de duplicatas.
B Padronização de datas (ISO).
C Enriquecimento de dados (adicionar CEP).
D Criação de novos pedidos de venda no sistema de origem.

Resposta Correta: D

O ETL apenas lê e transforma dados existentes. Ele nunca cria transações operacionais falsas no sistema de origem.

Questão 4: Modelagem

4. No Star Schema, o que é armazenado na Tabela Fato?

A Nomes de clientes e endereços.
B Métricas numéricas (vendas, quantidade) e chaves estrangeiras.
C Descrições de produtos.
D Logs de erro do servidor.

Questão 4: Modelagem

4. No Star Schema, o que é armazenado na Tabela Fato?

A Nomes de clientes e endereços.
B Métricas numéricas (vendas, quantidade) e chaves estrangeiras.
C Descrições de produtos.
D Logs de erro do servidor.

Resposta Correta: B

Fatos são números/eventos. Textos descritivos ficam nas Dimensões.

Questão 5: Integração

5. O que significa “Integração” no contexto de Data Warehousing?

A Padronizar dados de fontes heterogêneas para uma visão única.
B Conectar o computador na internet.
C Usar o Excel integrado com o Word.
D Fazer backup dos dados.

Questão 5: Integração

5. O que significa “Integração” no contexto de Data Warehousing?

A Padronizar dados de fontes heterogêneas para uma visão única.
B Conectar o computador na internet.
C Usar o Excel integrado com o Word.
D Fazer backup dos dados.

Resposta Correta: A

Integração é resolver conflitos (ex: Sexo M/F vs 0/1) para criar a Versão Única da Verdade.

Conclusão

O Data Warehouse não é apenas tecnologia, é **Estratégia**.

- Transforma o caos operacional em ordem analítica.
- Habilita a Inteligência Artificial e o BI.
- Garante a sobrevivência da empresa através do uso inteligente da informação.

Referências Bibliográficas I

- Ramez Elmasri and Shamkant B Navathe.
  Sistemas de Banco de Dados.
  Pearson, São Paulo, 7 edition, 2018.
  Referência base para conceitos de OLTP, SQL e normalização.

- William H Inmon.
  Building the Data Warehouse.
  John Wiley & Sons, Indianapolis, 4 edition, 2005.
  Referência clássica para a definição de DW (Orientado por assunto, Integrado, Não volátil, Variável no tempo).

- Ralph Kimball and Margy Ross.
  The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.
  John Wiley & Sons, Indianapolis, 3 edition, 2013.
  Referência principal para Modelagem Dimensional, Star Schema, Fatos e Dimensões.

Obrigado!

Dúvidas?