

# Big Data - Modelagem SBD OLTP

## Grupo

Alice Duarte Faria Ribeiro - DRE 122058907

Beatriz Farias do Nascimento – DRE 122053127

Gustavo do Amaral Roxo Pereira - DRE 122081146

## Link do github

<https://github.com/alicedfr/Big-Data-P2>

## Relatório da Parte II: Construção do Data Warehouse e Processo ETL

### 1. Introdução

A Parte II deste projeto teve como objetivo principal a **construção de um Modelo Dimensional de Data Warehouse (DW) e o desenvolvimento do processo de Extração, Transformação e Carga (ETL)** para integrar os dados transacionais de um grupo de empresas de locação de veículos. A finalidade é permitir a geração de relatórios gerenciais globais e a realização de análises de dados unificadas, superando os desafios da heterogeneidade dos sistemas operacionais independentes.

### 2. Arquitetura e Fluxo de Dados (ETL)

A solução foi estruturada em uma arquitetura de três camadas, que é o padrão da indústria para projetos de Business Intelligence.

#### 2.1. Camada 1: Staging Area (Área de Preparação)

- **Propósito:** Servir como repositório central para os dados brutos extraídos das quatro fontes de dados. Esta camada isola o Data Warehouse dos sistemas de origem, minimizando o impacto durante a extração.
- **Funcionalidade (**staging-area.sql**):** O script cria um conjunto de tabelas (**stg\_\***) projetadas para serem flexíveis o suficiente para receberem dados com diferentes nomes de colunas e formatos, um desafio comum na integração de sistemas legados.
- **Processo de Extração (**extracao.sql**):** Este script simula um processo de extração de um ambiente de produção real.
  - **Integração de Múltiplas Fontes:** Demonstra a lógica de conexão a quatro bancos de dados distintos.

- **Extração Incremental:** Implementa uma lógica de carga incremental (ex: `WHERE data_modificacao > @last_etl_run_timestamp`), garantindo que apenas os dados novos ou alterados sejam processados a cada ciclo do ETL. Isso é crucial para a performance e eficiência em ambientes com grandes volumes de dados.

## 2.2. Camada 2: Data Warehouse (Modelo Dimensional)

- **Propósito:** Armazenar os dados de forma otimizada para consultas analíticas e relatórios. É a "fonte única da verdade" para a gestão.
- **Funcionalidade (`data-warehouse.sql`):** O script implementa um **Esquema Estrela (Star Schema)**, que é a escolha ideal para este cenário devido à sua simplicidade e performance. O modelo é composto por:
  - **4 Dimensões Conformadas:** `dim_cliente`, `dim_veiculo`, `dim_patio` e `dim_tempo`. Elas fornecem o contexto descritivo ("quem, o quê, onde, quando") para os dados.
  - **3 Tabelas de Fatos:** `fato_locacoes`, `fato_reservas` e `fato_ocupacao_patio`. Elas armazenam as métricas quantitativas e os eventos de negócio.

## 2.3. Processos de Transformação e Carga (`transformacao.sql` e `carga.sql`)

- **Propósito:** Mover os dados da Staging Area para o Data Warehouse, aplicando as regras de negócio, limpeza e integração.
- **Funcionalidades (Transformação):**
  - **Limpeza e Padronização:** O script `transformacao.sql` é responsável por aplicar regras de negócio, como padronizar tipos de câmbio ou formatos de documento.
  - **Carga das Dimensões:** Ele popula as tabelas de dimensão, garantindo que não haja duplicatas de clientes, veículos ou pátios, criando assim uma visão mestre de cada entidade.
  - **Carga da Dimensão Tempo:** A criação de uma `procedure` para popular a `dim_tempo` é uma funcionalidade robusta que garante que a dimensão mais crítica do DW esteja sempre completa.
- **Funcionalidades (Carga):**
  - **Carga das Tabelas de Fatos:** O script `carga.sql` realiza a tarefa mais complexa: popular as tabelas de fatos. Ele faz o *lookup* (busca) das chaves primárias nas dimensões e calcula métricas derivadas, como `tempo_locacao` e `tempo_antecedencia`.
  - **Integridade Transacional:** A carga dos fatos é envolvida em uma `TRANSACTION`. Isso garante que o DW nunca fique em um estado

inconsistente; ou todas as cargas de fatos funcionam, ou nenhuma é aplicada.

### 3. Cobertura dos Requisitos de Negócio

A solução projetada atende diretamente a todos os requisitos de relatórios e análises definidos no contexto do projeto.

#### 3.1. Relatórios Gerenciais (`relatorios.sql`)

- **Controle de Pátio (Requisito 12.a):** O relatório é gerado a partir da `fato_ocupacao_patio`. Esta tabela foi projetada de forma inteligente como um fato transacional (registrando entradas e saídas), o que permite não apenas ver a ocupação em um dia, mas também analisar o fluxo de veículos ao longo do tempo.
- **Controle das Locações (Requisito 12.b):** O relatório utiliza a `fato_locacoes` para calcular métricas cruciais como a duração média das locações e, de forma inteligente, o tempo restante para devolução apenas para os contratos que estão `Ativos`.
- **Controle de Reservas (Requisito 12.c):** Utilizando a `fato_reservas`, o relatório fornece uma visão completa da demanda futura, cruzando os grupos de veículos desejados com os pátios de retirada e a origem dos clientes.
- **Grupos Mais Alugados (Requisito 12.d):** Este relatório cruza a `fato_locacoes` com as dimensões `dim_veiculo` e `dim_cliente` para identificar padrões de consumo e popularidade da frota por região.

#### 3.2. Análise Preditiva (Requisito 13)

- **Matriz de Movimentação para Cadeia de Markov:** A última consulta no script `relatorios.sql` é dedicada a esta tarefa. Ela calcula com precisão a matriz estocástica, mostrando o percentual de veículos que saem de um pátio de origem e são devolvidos em cada pátio de destino. Este é o input fundamental e necessário para a modelagem preditiva de ocupação de pátio usando Cadeias de Markov.

## 6. Conclusão

O modelo dimensional em estrela, juntamente com o processo ETL permite superar os desafios da heterogeneidade das fontes de dados, transformando dados operacionais brutos em informações estratégicas e acionáveis.

Os relatórios gerenciais e a capacidade de realizar análises avançadas, como a previsão de ocupação de pátio via Cadeia de Markov, são exemplos claros de como o DW pode apoiar a análise unificada e a tomada de decisões estratégicas.