# Descrição do Projeto

## Big Data - 2025.1

## Avaliação 02: Modelagem de Data Warehouse

## PARTE I - Modelagem SBD OLTP

Autores:

Augusto Fernandes Nodari DRE: 121131778
Henrique Almico Dias da Silva DRE: 124238228
João Pedro de Faria Sales DRE: 121056457
Vitor Rayol Taranto DRE: 121063585

## 1. Introdução

Este documento detalha o modelo dimensional estrela implementado para o Data Warehouse (DW) do consórcio de locadoras de veículos. O objetivo principal deste modelo é fornecer uma estrutura otimizada para consultas analíticas e relatórios gerenciais, consolidando dados de diversas fontes transacionais (OLTP) de maneira integrada e histórica, provenientes das seis empresas independentes que formam o consórcio. Serão abordadas as tabelas de dimensão e fato, suas ligações com as fontes de dados na camada de staging e a justificativa aprofundada para a inclusão de cada campo, sempre contextualizando com as necessidades específicas deste projeto.

## 2. Visão Geral do Modelo Dimensional Estrela

O Data Warehouse foi modelado seguindo o paradigma do esquema estrela, uma escolha estratégica para este projeto por diversas razões:

- **Simplicidade e Compreensibilidade:** Para um consórcio com múltiplas empresas e stakeholders com diferentes níveis de familiaridade com dados, a simplicidade do esquema estrela facilita a interpretação e o uso dos dados, permitindo que gestores de diferentes pátios ou unidades de negócio construam seus próprios relatórios com maior autonomia.
- **Performance de Consulta:** Com as dimensões desnormalizadas e as tabelas de fato contendo apenas chaves estrangeiras e medidas, as consultas analíticas, que frequentemente envolvem grandes volumes de dados, são executadas de forma muito mais eficiente, pois exigem menos `JOINs` complexos. Isso é crucial para dashboards e relatórios gerenciais que precisam de respostas rápidas.

- **Facilidade de Uso para BI:** Ferramentas de Business Intelligence (BI) se integram de forma mais fluida com esquemas estrela, tornando a construção de dashboards e relatórios uma tarefa mais intuitiva e menos suscetível a erros.

A ligação entre os sistemas OLTP legados de cada uma das seis empresas e o DW é realizada através de uma **camada de staging (`staging` schema)**. Nesta camada intermediária, os dados brutos são extraídos, passam por processos essenciais de padronização (resolvendo inconsistências de formato ou nomenclatura entre as diferentes empresas) e transformações iniciais antes de serem carregados nas tabelas de dimensão e fato do DW. Este passo é vital para garantir a unicidade e a integridade dos dados consolidados.

## 3. Tabelas de Dimensão (Schema `dw`)

As tabelas de dimensão fornecem o "quem", "o quê", "onde", "quando" e "como" do negócio, oferecendo o contexto descritivo para as medidas nas tabelas de fato.

### 3.1. `DimTempo`

- **Propósito:** Fornecer uma perspectiva temporal granular e consistente para todas as análises no DW, permitindo a agregação de dados por diferentes níveis de tempo (dia, mês, ano, trimestre, semestre) de forma unificada para todas as empresas do consórcio.
- **Ligação com Fontes de Dados:** É populada independentemente dos dados transacionais, através de uma `generate_series` de datas (no script `load.sql`/`transform.sql`), geralmente cobrindo um período amplo (e.g., de 2020 a 2050). Isso garante que todas as datas em um intervalo predefinido estejam disponíveis para análise, mesmo que não haja transações específicas em um determinado dia. As datas de transação da camada de staging (`stg_reservas.data_reserva`, `stg_locacoes.data_retirada_real`, `stg_locacoes.data_devolucao_real`, etc.) são usadas para buscar a `sk_tempo` correspondente.
- **Campos Justificativa:**
    - `sk_tempo` (SERIAL PRIMARY KEY): **Chave surrogada**, um número inteiro sequencial, otimizada para joins de alta performance e completamente independente de chaves naturais dos sistemas fonte. Essencial para a integridade referencial do DW.
    - `data_completa` (DATE NOT NULL UNIQUE): A data exata do dia. Serve como a base única para os demais atributos temporais, sendo o ponto de conexão direto para associar qualquer evento de locação ou reserva a um momento específico no tempo.
    - `ano` (INT NOT NULL): O ano da `data_completa`. Permite análises anuais de performance consolidada, como "total de locações do consórcio por ano" ou "crescimento da receita anual por pátio".

- `mes` (INT NOT NULL): O número do mês (1-12). Útil para análises mensais e comparações de desempenho entre meses ao longo dos anos, revelando sazonalidades para o planejamento de frota.
- `dia` (INT NOT NULL): O dia do mês (1-31). Permite análises diárias e de tendências de curtíssimo prazo, auxiliando na gestão operacional dos pátios.
- `trimestre` (INT NOT NULL): O trimestre do ano (1-4). Agrupamento para relatórios trimestrais de performance, alinhando com ciclos de planejamento de negócios.
- `semestre` (INT NOT NULL): O semestre do ano (1-2). Agrupamento para relatórios semestrais de performance, útil para avaliações de médio prazo.
- `dia_da_semana` (INT NOT NULL): O dia da semana (1=Segunda, 7=Domingo). Permite análises por dia da semana, identificando padrões de demanda que podem otimizar a disponibilidade de veículos (ex: maior demanda por SUVs aos finais de semana).
- `nome_mes` (VARCHAR(20) NOT NULL): Nome do mês por extenso (ex: 'Janeiro'). Facilita a leitura e apresentação dos relatórios para usuários de negócio.
- `nome_dia_semana` (VARCHAR(20) NOT NULL): Nome do dia da semana por extenso (ex: 'Domingo'). Melhora a legibilidade e a inteligibilidade dos relatórios.
- `feriado` (BOOLEAN DEFAULT FALSE): Indica se a data é um feriado. Essencial para analisar o impacto de feriados na demanda por locações/reservas e no fluxo de veículos entre os pátios, ajudando no planejamento de equipes e frota.

### 3.2. `DimCliente`

- **Propósito:** Fornecer informações contextuais sobre os clientes de todas as locadoras associadas, permitindo análises demográficas, de comportamento de consumo e de fidelidade. Implementa o **Slowly Changing Dimension (SCD) Tipo 2** para preservar o histórico completo de alterações em atributos-chave do cliente.
- **Ligação com Fontes de Dados:** Populada a partir de `staging.stg_clientes` (conforme `load.sql`/`transform.sql`). A lógica de SCD Tipo 2 no ETL detecta mudanças em atributos como `nome_completo`, `cidade` ou `estado` do cliente. Se um cliente que morava no "Rio de Janeiro" muda para "Niterói", o registro antigo (Rio) é encerrado (`data_fim` preenchida, `versao_atual = FALSE`) e um novo registro para o mesmo `cliente_id` é criado (Niterói), com nova `data_inicio` e `versao_atual = TRUE`. Isso garante que relatórios históricos de locações reflitam a cidade do cliente no momento da transação.
- **Campos Justificativa:**

- `sk_cliente` (SERIAL PRIMARY KEY): **Chave surrogada** para gerenciar o histórico de clientes (SCD Tipo 2) e otimizar joins. Cada alteração relevante no perfil do cliente gera um novo `sk_cliente`.
- `cliente_id` (INT NOT NULL): **Chave natural** do cliente do sistema OLTP. Mantida para rastrear o cliente original e vincular diferentes versões históricas do mesmo cliente.
- `nome_completo` (VARCHAR(255) NOT NULL): Nome completo do cliente. Fundamental para identificação, personalização e segmentação em campanhas de marketing ou relatórios de atendimento.
- `tipo_pessoa` (CHAR(1) NOT NULL): Indica se é Pessoa Física ('F') ou Pessoa Jurídica ('J'). Permite segmentar análises por tipo de cliente, identificando, por exemplo, a demanda por frotas corporativas versus aluguéis individuais.
- `cidade` (VARCHAR(100)): Cidade de residência do cliente. Crucial para o relatório "Grupos Mais Alugados" cruzando com a origem do cliente, como "clientes de São Paulo que alugam SUVs no Rio".
- `estado` (VARCHAR(50)): Estado de residência do cliente. Complementa a `cidade` para análises geográficas mais amplas e distribuição de clientes pelo país.
- `data_inicio` (TIMESTAMP NOT NULL): Data a partir da qual esta versão específica do registro do cliente é válida. Essencial para a funcionalidade do SCD Tipo 2.
- `data_fim` (TIMESTAMP): Data até a qual esta versão do registro do cliente foi válida. Será nulo se for a versão atualmente ativa. Essencial para a funcionalidade do SCD Tipo 2.
- `versao_atual` (BOOLEAN DEFAULT TRUE): Indica se este é o registro mais recente e ativo do cliente. Facilita a consulta pela versão vigente do cliente para relatórios que não exigem histórico completo.

### 3.3. `DimVeiculo`

- **Propósito:** Descrever de forma abrangente os atributos de cada veículo da frota consolidada do consórcio, permitindo análises sobre o desempenho de diferentes tipos de veículos, a composição da frota e a popularidade de modelos específicos.
- **Ligação com Fontes de Dados:** Populada a partir de `staging.stg_veiculos`. Para enriquecer a dimensão, é realizado um `JOIN` com `staging.stg_grupos_veiculos` para trazer as informações do grupo (`nome_grupo`, `descricao`) diretamente para esta dimensão. Esta desnormalização (vista no `load.sql`/`transform.sql`) otimiza o desempenho de consultas, pois a maioria dos relatórios de frota e locação precisará de informações do grupo.
- **Campos Justificativa:**

- `sk_veiculo` (SERIAL PRIMARY KEY): **Chave surrogada**, otimiza joins em grandes tabelas de fatos e é independente de chaves naturais, permitindo flexibilidade se as chaves OLTP mudarem.
- `veiculo_id` (INT NOT NULL): **Chave natural** do veículo do sistema OLTP. Mantida para rastrear o veículo original e para fins de auditoria ou integração com outros sistemas operacionais.
- `placa` (VARCHAR(10) NOT NULL): Placa de identificação única do veículo. Essencial para rastreamento individual e conformidade legal.
- `marca` (VARCHAR(50) NOT NULL): Marca do veículo (ex: 'Chevrolet', 'Volkswagen'). Permite análises de popularidade e desempenho por marca em toda a frota do consórcio.
- `modelo` (VARCHAR(50) NOT NULL): Modelo do veículo (ex: 'Onix', 'T-Cross'). Permite análises mais detalhadas por modelo, identificando os carros mais alugados ou os que mais necessitam de manutenção.
- `cor` (VARCHAR(30) NOT NULL): Cor do veículo. Pode ser um atributo de interesse para algumas análises específicas ou segmentação visual.
- `ano_fabricacao` (INT NOT NULL): Ano de fabricação do veículo. Útil para análises de frota por idade, auxiliando na decisão de renovação e depreciação.
- `mecanizacao` (VARCHAR(20) NOT NULL): Tipo de mecanização (ex: 'Automático', 'Manual'). Atributo importante para o "Controle de Pátio" e para entender a demanda por diferentes tipos de câmbio.
- `grupo_nome` (VARCHAR(50) NOT NULL): Nome do grupo ao qual o veículo pertence (ex: 'Econômico', 'SUV', 'Luxo'). Este campo desnormalizado é crucial para o "Controle de Pátio" e "Grupos Mais Alugados", permitindo a agregação por categoria de veículo em toda a frota do consórcio.
- `grupo_descricao` (TEXT): Descrição detalhada do grupo do veículo. Desnormalizado para fornecer contexto adicional sobre a categoria do veículo sem a necessidade de um join extra.

### 3.4. `DimPatio`

- **Propósito:** Fornecer informações claras e unificadas sobre os seis pátios do consórcio, que funcionam como locais de retirada e devolução de veículos. Essencial para análises espaciais, de movimentação da frota e para a previsão de ocupação de cada unidade.
- **Ligação com Fontes de Dados:** Populada a partir de `staging.stg_patios`. O ETL garante que os pátios de todas as empresas associadas sejam consolidados aqui.
- **Campos Justificativa:**
  - `sk_patio` (SERIAL PRIMARY KEY): **Chave surrogada**, otimiza joins e é fundamental para a Cadeia de Markov e análises de fluxo de veículos.

- `patio_id` (INT NOT NULL): **Chave natural** do pátio do sistema OLTP. Mantida para rastreamento e auditoria.
- `nome_patio` (VARCHAR(100) NOT NULL): Nome do pátio (ex: 'Aeroporto do Galeão', 'Santos Dumont', 'Rodoviária', 'Shopping Rio Sul', 'Nova América', 'Barra Shopping'). Essencial para identificar os locais nos relatórios gerenciais e para que os usuários de negócio compreendam o contexto geográfico.
- `endereco_patio` (VARCHAR(255) NOT NULL): Endereço completo do pátio. Fornece contexto geográfico adicional, útil para visualizações em mapas ou para entender a proximidade entre pátios.

### 3.5. `DimGrupoVeiculo`

- **Propósito:** Categorizar os veículos em grupos lógicos para fins de reserva e análise, refletindo as ofertas de serviço padronizadas do consórcio de locadoras.
- **Ligação com Fontes de Dados:** Populada a partir de `staging.stg_grupos_veiculos`.
- **Campos Justificativa:**
  - `sk_grupo_veiculo` (SERIAL PRIMARY KEY): **Chave surrogada**, otimiza joins.
  - `grupo_id` (INT NOT NULL): **Chave natural** do grupo do sistema OLTP. Mantida para rastreamento.
  - `nome_grupo` (VARCHAR(50) NOT NULL): Nome do grupo de veículo (ex: 'SUV', 'Econômico', 'Luxo'). Este é o principal atributo para análises por categoria de veículo, tanto em reservas (o cliente reserva um grupo) quanto em locações (o veículo efetivamente alugado pertence a um grupo).
  - `valor_diaria_base` (DECIMAL(10, 2) NOT NULL): Valor da diária base para veículos neste grupo. Útil para análises de precificação, projeção de receita e comparação de desempenho financeiro entre diferentes grupos de veículos.

## 4. Tabelas de Fato (Schema `dw`)

As tabelas de fato capturam os eventos e transações do negócio, juntamente com as medidas numéricas associadas, formando o centro do esquema estrela.

### 4.1. `FatoLocacoes`

- **Propósito:** Registrar as informações detalhadas de cada locação de veículo concretizada em qualquer um dos pátios do consórcio, permitindo análises de desempenho financeiro, eficiência operacional, duração das locações e, crucialmente, a movimentação da frota entre os pátios.

- **Grão:** Uma linha por locação de veículo. Este é o nível mais atômico de detalhe para transações de aluguel.
- **Ligação com Fontes de Dados:** Populada a partir de `staging.stg_locacoes`. O ETL realiza `JOINs` com as dimensões (`DimTempo`, `DimCliente`, `DimVeiculo`, `DimPatio`) usando as chaves naturais e buscando as `sk_` correspondentes. A lógica de `ON CONFLICT DO UPDATE SET` no `load.sql`/`transform.sql` é vital: permite que uma locação iniciada seja atualizada posteriormente com a `data_devolucao_real` e `valor_total_final` quando o veículo é devolvido, refletindo a conclusão da transação.
- **Campos Justificativa:**
    - `locacao_id` (INT PRIMARY KEY): **Chave natural** da locação do sistema OLTP. Usada como chave primária na fato para garantir a unicidade de cada transação de aluguel no DW.
    - `sk_data_retirada` (INT NOT NULL): **Chave estrangeira** para `DimTempo`, indicando a data de retirada real do veículo. Permite análises temporais relacionadas ao início da locação, como o número de locações por dia/mês/ano.
    - `sk_data_devolucao` (INT): **Chave estrangeira** para `DimTempo`, indicando a data de devolução real do veículo. Pode ser nula para locações que ainda estão em andamento. Essencial para calcular a duração real da locação e para a análise de Cadeia de Markov.
    - `sk_cliente` (INT NOT NULL): **Chave estrangeira** para `DimCliente`, conectando a locação ao cliente que a realizou. Permite análises de comportamento do cliente, como "quantas locações um cliente específico realizou" ou "quais grupos de veículos são mais populares entre clientes de determinada cidade".
    - `sk_veiculo` (INT NOT NULL): **Chave estrangeira** para `DimVeiculo`, conectando a locação ao veículo específico alugado. Permite análises de desempenho individual do veículo (ex: "quantas vezes um carro foi alugado") e da frota (ex: "utilização média dos veículos SUV").
    - `sk_patio_retirada` (INT NOT NULL): **Chave estrangeira** para `DimPatio`, indicando o pátio onde o veículo foi retirado. Crucial para o "Controle de Pátio" e, principalmente, para a construção da matriz de transição da Cadeia de Markov (origem da movimentação).
    - `sk_patio_devolucao` (INT): **Chave estrangeira** para `DimPatio`, indicando o pátio onde o veículo foi devolvido. Pode ser nula para locações ativas. Essencial para a análise de Cadeia de Markov (destino da movimentação) e para entender o fluxo de veículos entre os diferentes pátios do consórcio.
    - `valor_total_previsto` (DECIMAL(10, 2) NOT NULL): O valor total da locação previsto no momento da retirada. Medida para análise de metas

financeiras e para comparar com o valor final, identificando desvios de precificação.
    - `valor_total_final` (DECIMAL(10, 2)): O valor total final da locação, considerando quaisquer ajustes (ex: acessórios adicionais, atrasos) na devolução. A medida real de receita gerada por locação.
    - `dias_locacao_previstos` (INT): Número de dias previstos para a locação. Medida calculada (data prevista de devolução - data de retirada) para fins de planejamento e comparação com a duração real.
    - `dias_locacao_reais` (INT): Número de dias reais da locação. Medida calculada (data de devolução real - data de retirada real) para análise de performance e utilização.
    - `quantidade_locacoes` (INT DEFAULT 1): Uma medida aditiva constante de 1. Usada para contar o número de locações em agregações (ex: "total de locações no pátio do Galeão").

### 4.2. `FatoReservas`

- **Propósito:** Registrar as intenções de aluguel (reservas) feitas pelos clientes em qualquer um dos pátios, permitindo a análise da demanda futura por grupos de veículos em pátios específicos e a previsão de ocupação com base nas reservas.
- **Grão:** Uma linha por reserva de veículo.
- **Ligação com Fontes de Dados:** Populada a partir de `staging.stg_reservas`. O ETL realiza `JOINs` com as dimensões (`DimTempo`, `DimCliente`, `DimGrupoVeiculo`, `DimPatio`). A inclusão de `reserva_id` no `ON CONFLICT DO NOTHING` no `load.sql`/`transform.sql` garante que as reservas sejam inseridas uma única vez.
- **Campos Justificativa:**
    - `reserva_id` (INT PRIMARY KEY): **Chave natural** da reserva do sistema OLTP. Usada como chave primária na fato.
    - `sk_data_reserva` (INT NOT NULL): **Chave estrangeira** para `DimTempo`, indicando a data em que a reserva foi efetivada. Permite analisar tendências de reservas ao longo do tempo (ex: "quando as reservas são mais frequentemente feitas?").
    - `sk_data_prevista_retirada` (INT NOT NULL): **Chave estrangeira** para `DimTempo`, indicando a data prevista de retirada do veículo. Essencial para o "Controle de Reservas" e para prever a demanda futura (ex: "quantas reservas para SUVs no pátio Santos Dumont para a próxima semana?").
    - `sk_cliente` (INT NOT NULL): **Chave estrangeira** para `DimCliente`, conectando a reserva ao cliente. Permite análises de perfil de cliente que faz reservas (ex: "clientes PJ tendem a reservar com mais antecedência?").
    - `sk_grupo_veiculo` (INT NOT NULL): **Chave estrangeira** para `DimGrupoVeiculo`, indicando o tipo de veículo que foi reservado.

Essencial para entender a demanda futura por categoria (ex: "qual grupo de veículo está mais reservado para o próximo feriado?").
    - `sk_patio_retirada` (INT NOT NULL): **Chave estrangeira** para `DimPatio`, indicando o pátio de retirada desejado pelo cliente. Crucial para a alocação de frota futura e para a análise de demanda por pátio.
    - `status_reserva` (VARCHAR(20)): O status atual da reserva (ex: 'Ativa', 'Cancelada', 'Concluída'). Permite filtrar e analisar reservas por seu estado, focando apenas nas reservas confirmadas ou para entender as taxas de cancelamento.
    - `quantidade_reservas` (INT DEFAULT 1): Uma medida aditiva constante de 1. Usada para contar o número de reservas em agregações (ex: "total de reservas no pátio do Barra Shopping").

## 5. Conclusão

O modelo dimensional estrela desenvolvido para o Data Warehouse oferece uma solução robusta e altamente eficaz para os desafios de Business Intelligence enfrentados pelo consórcio de locadoras de veículos. Ao desnormalizar atributos descritivos em dimensões ricas e centralizar as medidas em tabelas de fatos granulares, o modelo atinge os seguintes benefícios cruciais, diretamente alinhados aos objetivos do projeto:

- **Visão Unificada e Global da Operação:** O DW proporciona, de fato, uma visão integrada e unificada dos dados operacionais das seis empresas distintas. Isso supera o desafio inicial da heterogeneidade de sistemas e permite que a gestão do consórcio tenha uma perspectiva consolidada e estratégica, essencial para decisões coordenadas e para otimizar recursos em nível de grupo.
- **Suporte Abrangente à Decisão Estratégica e Operacional:** A estrutura dimensional permite a geração eficiente e rápida de todos os relatórios gerenciais globais solicitados. Por exemplo, o "Controle de Pátio" pode agora comparar a ocupação de veículos SUV no Aeroporto do Galeão vs. Santos Dumont em tempo real; o relatório de "Grupos Mais Alugados" pode revelar que, globalmente, clientes do sul do país preferem veículos de luxo para aluguel no Rio, informando decisões sobre realocação de frota.
- **Capacidade de Análises Avançadas e Preditivas:** A granularidade da `FatoLocacoes`, especialmente com as dimensões de pátio de retirada e devolução, fornece a base perfeita para a construção da matriz estocástica necessária para a análise da Cadeia de Markov. Isso capacita o consórcio a prever a ocupação futura de cada pátio, permitindo a **otimização logística** (ex: realocar veículos proativamente de um pátio com excesso para outro com demanda futura esperada) e a **minimização de custos** (reduzindo ociosidade ou a falta de frota em momentos de pico).
- **História Completa e Evolução de Negócio:** A implementação de Slowly Changing Dimensions (SCD Tipo 2) para a `DimCliente` garante que o histórico de mudanças

nos dados do cliente (e.g., mudanças de endereço ou tipo de pessoa jurídica) seja preservado. Isso permite análises precisas de comportamentos passados do cliente, mesmo que seus dados atuais tenham mudado. A modularidade do esquema também prepara o DW para futuras expansões, como a inclusão de novas medidas financeiras ou dimensões de marketing, sem a necessidade de grandes reestruturações.

Em resumo, o modelo dimensional estrela proposto não apenas atende aos requisitos atuais de relatórios e análises, mas também estabelece uma fundação escalável e eficiente para futuras iniciativas de Business Intelligence. Isso permitirá que as locadoras associadas, agindo como um consórcio unificado, tomem decisões muito mais informadas e estratégicas, resultando em maior eficiência operacional e maximização do retorno sobre o investimento.