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 apresenta um relatório detalhado sobre o desenvolvimento do processo de Extração, Transformação e Carga (ETL) e a modelagem do Data Warehouse (DW) para um consórcio de seis empresas independentes de aluguel de automóveis. O objetivo principal é unificar os dados operacionais de clientes, frota, pátios, reservas e locações de diferentes sistemas transacionais (OLTP) em um ambiente analítico coeso. Este DW permitirá a geração de relatórios gerenciais globais e análises unificadas, como a previsão de ocupação de pátios, essencial para a otimização das operações conjuntas.

2. Escolhas e Observações no Processo ETL

O processo ETL foi projetado para consolidar dados de fontes heterogêneas, provenientes dos sistemas OLTP de cada empresa associada, em um esquema estrela no Data Warehouse. As principais escolhas e observações durante o desenvolvimento são:

2.1. Camada de Staging

Foi implementada uma camada de staging (`staging` schema) como um intermediário essencial entre os sistemas OLTP e o DW. Esta camada serve para:

- **Padronização:** Receber os dados brutos de cada empresa, garantindo que atributos como `data_cadastro` (em `stg_clientes`) e `data_prevista_retirada` (em `stg_reservas`) sejam consistentemente formatados como `TIMESTAMP WITH TIME ZONE`, independentemente da granularidade ou formato original nos sistemas OLTP. Isso evita problemas de compatibilidade de tipos de dados durante a

carga no DW e assegura que todas as datas e horas possam ser comparadas e agregadas corretamente.

- **Limpeza e Transformação:** Realizar operações iniciais de limpeza, como o tratamento de valores nulos, preenchendo-os com valores padrão ou marcadores específicos (e.g., `'N/A'` para descrições vazias). Transformações simples podem incluir a conversão de unidades, a derivação de novos atributos (e.g., `dias_locacao_previstos` a partir de `data_retirada_real` e `data_devolucao_prevista` em `stg_locacoes`), ou a padronização de campos de texto, como a capitalização de nomes de cidades e estados. Embora o SQL fornecido para staging apenas crie as tabelas, em um ambiente real, esta etapa incluiria scripts para carregar os dados brutos e aplicar transformações básicas antes da carga para o DW.
- **Identificação da Fonte (`source_company_id`):** A coluna `source_company_id` foi adicionada a todas as tabelas de staging (e posteriormente, inferida em dimensões e fatos relevantes do DW, como `DimVeiculo` para a origem da frota). Esta é uma escolha crucial para rastrear a origem dos dados de cada uma das seis empresas (e.g., 'Galeão', 'Santos Dumont', 'Rodoviária', etc.). Isso permite análises por "origem" de frota ou cliente, conforme solicitado nos relatórios gerenciais (e.g., "Controle de pátio: quantitativo de veículos no pátio por 'grupo' e 'origem'").
- **Timestamp de Extração (`extraction_date`):** A coluna `extraction_date` registra o momento exato em que os dados foram extraídos do sistema OLTP para a camada de staging. Este campo é vital para o controle de cargas incrementais (identificando quais registros foram adicionados ou modificados desde a última extração) e para o rastreamento da frescura dos dados, garantindo que as análises sempre utilizem as informações mais recentes disponíveis.

2.2. Extração e Carga (E & L)

Assumiu-se que a extração dos dados dos sistemas OLTP para as tabelas de staging seria feita periodicamente, por exemplo, diariamente ao final do dia ou em janelas de manutenção, para minimizar o impacto nos sistemas transacionais. A carga da camada de staging para o DW seria então incremental, focando apenas nos dados novos ou alterados para otimizar o desempenho e reduzir o tempo de processamento. Estratégias como a detecção de mudanças (Change Data Capture - CDC) ou a comparação de `extraction_date` seriam empregadas para identificar e carregar apenas os registros modificados ou inseridos. Para fins deste projeto, os scripts de ETL focam na estrutura do DW, e a lógica de `INSERT INTO ... SELECT FROM` seria implementada para popular as tabelas de fato e dimensão, orquestrada por uma ferramenta ETL.

2.3. Tratamento de Chaves Surrogadas

Uma observação importante é a geração de chaves surrogadas (`sk_`) para todas as dimensões no DW (ex: `sk_tempo`, `sk_cliente`, `sk_veiculo`). Esta é uma prática padrão e essencial em modelagem dimensional por diversas razões:

- **Independência da Fonte:** Desacopla as chaves primárias do DW das chaves naturais (originais) dos sistemas OLTP. Isso é fundamental em um ambiente multi-empresa onde chaves naturais podem não ser únicas globalmente.
- **Gerenciamento de SCDs:** Facilita a implementação de Slowly Changing Dimensions (SCDs), como o Tipo 2 em `DimCliente`, onde um novo `sk_cliente` é gerado a cada alteração de um atributo rastreado, preservando o histórico completo do cliente.
- **Desempenho:** Chaves surrogadas são tipicamente números inteiros sequenciais, otimizando o desempenho de junções (JOINs) entre tabelas de fatos e dimensões, que são operações frequentes em consultas analíticas.

2.4. Problemas e Observações Gerais no Desenvolvimento

Durante o processo de desenvolvimento e modelagem, nós enfrentamos alguns pontos cruciais e desafios que foram observados:

- **Consistência de Dados entre Fontes:** Um problema inerente ao ambiente multi-empresa foi a variabilidade na qualidade e consistência dos dados provenientes de diferentes sistemas OLTP. Embora a camada de staging tenha mitigado parte disso, garantir a uniformidade completa (e.g., formatos de endereço, categorias de veículos ligeiramente diferentes) exigiu de nós a criação de regras de negócio ETL mais complexas e validações robustas. A padronização foi uma escolha consciente para evitar problemas a jusante.
- **Granularidade da DimTempo:** A escolha de uma granularidade diária para `DimTempo` foi adequada para a maioria dos relatórios solicitados. No entanto, para análises de ocupação de pátio em tempo real ou previsão de curtíssimo prazo, percebemos que uma granularidade mais fina (hora, minuto) poderia ser necessária, o que implicaria em um volume de dados significativamente maior e complexidade adicional no ETL. Optamos pela granularidade diária para equilibrar performance e requisitos.
- **Gestão de Dados Incompletos/Nulos:** A identificação e o tratamento de dados incompletos ou nulos nos sistemas OLTP de origem representaram um desafio contínuo para nós. Fizemos escolhas para preencher nulos com valores padrão (e.g., 'Desconhecido') onde apropriado, ou para permitir nulos nas tabelas de staging/DW se a informação não fosse crítica para as análises.
- **Desacoplamento de Fatos e Dimensões:** A modelagem dimensional com esquema estrela foi uma escolha deliberada que fizemos para otimizar o desempenho das consultas analíticas em detrimento de uma normalização extrema, que seria mais adequada para um sistema transacional. Isso simplificou as consultas para os

usuários e as ferramentas de BI, mas exigiu de nós um planejamento cuidadoso das dimensões e dos atributos a serem desnormalizados.

- **Complexidade do ETL vs. Manutenibilidade:** O balanço entre a complexidade da lógica ETL necessária para transformar os dados e a facilidade de manutenção futura foi uma consideração constante para o nosso grupo. Priorizamos a clareza e a modularidade dos scripts, mesmo que isso implicasse em mais etapas, para facilitar a depuração e futuras modificações.
- **Evolução do Schema:** Prevemos que o schema do DW pode precisar evoluir com novas necessidades de negócio. A utilização de chaves surrogadas e a separação clara entre as camadas de staging e DW facilitam essa evolução, minimizando o impacto em outras partes do sistema.

3. Modelo Dimensional do Data Warehouse

O modelo dimensional foi projetado seguindo o esquema estrela, com tabelas de fatos centralizadas e tabelas de dimensão desnormalizadas. As escolhas feitas visam otimizar o desempenho de consultas analíticas e a compreensibilidade para usuários de negócio, facilitando a navegação e a exploração dos dados.

3.1. Dimensões Implementadas (`dw` schema)

- **`DimTempo`:** Essencial para todas as análises temporais, contendo atributos detalhados como `ano`, `mes`, `dia`, `trimestre`, `semestre`, `dia_da_semana`, `nome_mes`, `nome_dia_semana` e `feriado`. A chave `sk_tempo` é baseada na `data_completa`, permitindo que os usuários analisem tendências de locação por período, identifiquem picos em feriados ou comparem o desempenho mensal e trimestral.
- **`DimCliente`:** Contém informações sobre os clientes (`nome_completo`, `tipo_pessoa`, `cidade`, `estado`). A inclusão de `data_inicio`, `data_fim` e `versao_atual` indica a implementação de Slowly Changing Dimensions do Tipo 2. Isso significa que se um cliente muda de cidade, o registro antigo é mantido (com `versao_atual = FALSE` e `data_fim` preenchida), e um novo registro é criado (com `versao_atual = TRUE`). Isso é crucial para relatórios que precisam de um "estado" do cliente em um determinado momento histórico (e.g., "Quais os grupos de veículos mais alugados, cruzando, eventualmente, com a origem dos clientes" considerando a cidade do cliente no momento da locação).
- **`DimVeiculo`:** Detalha os veículos da frota, incluindo `placa`, `marca`, `modelo`, `cor`, `ano_fabricacao`, `mecanizacao` e atributos do grupo a que pertence (`grupo_nome` e `grupo_descricao`). A desnormalização do `grupo_nome` e `grupo_descricao` diretamente nesta dimensão evita joins adicionais com `DimGrupoVeiculo` para muitas consultas que envolvem atributos do veículo e seu grupo, otimizando o desempenho. Esta dimensão permite filtrar e agrupar veículos

para análises de frota, como as do "Controle de Pátio" por marca, modelo e tipo de mecanização.

- **`DimPatio`:** Contém informações sobre os pátios (`nome_patio` e `endereco_patio`). A chave `sk_patio` permite identificar de forma única cada localidade física. Esta dimensão é crucial para análises espaciais e para entender a movimentação dos veículos entre os pátios, como na previsão de ocupação.
- **`DimGrupoVeiculo`:** Informações sobre os grupos de veículos (ex: 'Econômico', 'SUV', 'Luxo'), com `nome_grupo` e `valor_diaria_base`. Esta dimensão é útil para análises de reservas (onde a reserva é feita para um grupo, não um veículo específico) e para agrupar veículos por categoria em relatórios de desempenho e popularidade.

3.2. Tabelas de Fato

Foram modeladas duas tabelas de fatos principais para atender aos requisitos de relatórios e análises, cada uma com sua granularidade específica e métricas relevantes:

- **`FatoLocacoes`:**
  - **Grão:** Cada linha representa uma locação individual concretizada. Este é o nível mais atômico de detalhe para transações de aluguel.
  - **Medidas:** `valor_total_previsto` (para análise de metas), `valor_total_final` (para receita real), `dias_locacao_previstos` (para planejamento), `dias_locacao_reais` (para duração efetiva), e `quantidade_locacoes` (sempre 1, mas útil para somar o número total de locações).
  - **Dimensões Conectadas:** `DimTempo` (duas vezes, para `sk_data_retirada` e `sk_data_devolucao`), `DimCliente`, `DimVeiculo`, `DimPatio` (duas vezes, para `sk_patio_retirada` e `sk_patio_devolucao`).
  - **Observações:** Esta tabela é fundamental para o "Controle das Locações", permitindo calcular o tempo médio de aluguel, o valor total das locações e a contagem de veículos alugados. Mais importante, ela é a fonte de dados principal para a análise de Cadeia de Markov, pois os pares `sk_patio_retirada` e `sk_patio_devolucao` permitem rastrear a movimentação exata dos veículos entre os pátios.
- **`FatoReservas`:**
  - **Grão:** Cada linha representa uma reserva individual efetuada por um cliente.
  - **Medidas:** `quantidade_reservas` (sempre 1, permitindo a contagem total de reservas).
  - **Dimensões Conectadas:** `DimTempo` (para `sk_data_reserva` e `sk_data_prevista_retirada`), `DimCliente`, `DimGrupoVeiculo`, `DimPatio` (para `sk_patio_retirada`).

    - **Observações:** Essencial para o "Controle de Reservas", permitindo analisar a demanda futura por grupo de veículo, pátio de retirada, e tempo até a retirada. A conexão com `DimGrupoVeiculo` é mais apropriada aqui, pois as reservas são feitas para uma categoria de veículo e não para um veículo específico.

4. Relatórios Gerenciais e Análise de Previsão

O modelo de DW suporta diretamente os relatórios gerenciais globais e a análise de previsão de ocupação de pátio, conforme detalhado abaixo:

- **Controle de Pátio (12.a):** Este relatório, que visa quantificar veículos por "grupo", "origem", marca, modelo e tipo de mecanização, é facilmente derivado da `FatoLocacoes` (para veículos que estão atualmente locados) e dos dados de frota armazenados na `DimVeiculo` (para veículos disponíveis ou em manutenção). A `source_company_id` (que será propagada para a `DimVeiculo` durante o ETL ou inferida através do `veiculo_id` e `source_company_id` da staging) é crucial para a dimensão "origem", permitindo diferenciar a frota da empresa dona do pátio da frota de outras empresas associadas. Agrupamentos por `marca`, `modelo` e `mecanizacao` são diretamente possíveis pelos atributos em `DimVeiculo`.
- **Controle das Locações (12.b):** O quantitativo de veículos alugados por "grupo", o tempo de locação e o tempo restante para devolução são calculados a partir da `FatoLocacoes`. As medidas `dias_locacao_previstos` e `dias_locacao_reais` (quando a `data_devolucao_real` é preenchida) permitem a análise da duração das locações. Para o "tempo restante para devolução", uma consulta na `FatoLocacoes` pode cruzar a `data_devolucao_prevista` (da `DimTempo` associada) com a data atual, filtrando por locações ainda ativas (onde `data_devolucao_real` é nula).
- **Controle de Reservas (12.c):** Este relatório, que mede quantas reservas por "grupo" de veículo, "pátio" de retirada, tempo de retirada futura e cidades de origem dos clientes, é construído a partir da `FatoReservas`. A `DimGrupoVeiculo` permite agrupar por grupo, `DimPatio` por pátio de retirada, e `DimCliente` (com `cidade`) pela origem do cliente. A `DimTempo` conectada à `sk_data_prevista_retirada` permite calcular o "tempo de retirada futura" (e.g., reservas para a próxima semana, mês).
- **Grupos Mais Alugados (12.d):** Para identificar os "grupos" de veículos mais alugados, possivelmente cruzando com a origem dos clientes, a `FatoLocacoes` é a tabela de fato principal. Agrupando por `DimGrupoVeiculo` (`nome_grupo`) e `DimCliente` (`cidade`), e somando a `quantidade_locacoes`, é possível obter insights sobre a popularidade de certos grupos de veículos em diferentes localidades de origem dos clientes.

- **Previsão de Ocupação de Pátio (Cadeia de Markov - 13):** A `FatoLocacoes`, com suas chaves de pátio de retirada (`sk_patio_retirada`) e devolução (`sk_patio_devolucao`), é a fonte de dados perfeita para construir a matriz estocástica necessária para o modelo de Cadeia de Markov. Ao contar a frequência de transições entre cada par (pátio de retirada, pátio de devolução) na `FatoLocacoes`, é possível calcular os percentuais de movimentação da frota entre os pátios. Isso permite responder, para cada pátio, qual o percentual de veículos que retorna ao mesmo pátio de onde foi retirado e o percentual que é entregue em cada um dos outros pátios, alimentando diretamente o modelo preditivo.

5. Conclusão

O modelo de Data Warehouse desenvolvido, em conjunto com o processo ETL proposto, oferece uma solução robusta, flexível e escalável para as necessidades analíticas das empresas associadas de aluguel de veículos. Em vistas dos objetivos que nortearam a necessidade por uma solução de Data Warehouse, os resultados obtidos são altamente satisfatórios e tangíveis:

- **Visão Unificada e Global:** O DW proporciona, de fato, uma visão integrada e unificada dos dados operacionais de seis empresas distintas, superando o desafio inicial da heterogeneidade de sistemas. Isso permite que a gestão do consórcio tenha uma perspectiva global em vez de silos de informação, essencial para decisões coordenadas.
- **Suporte à Decisão Estratégica e Operacional:** A estrutura dimensional permite a geração eficiente e rápida dos relatórios gerenciais globais solicitados, empoderando os gestores com informações precisas e acionáveis sobre a frota, o desempenho das locações, a demanda por reservas e a ocupação dos pátios. Por exemplo, a capacidade de identificar que "muitos clientes de São Paulo alugam SUVs no Rio" permite que a empresa reforce esse grupo nos pátios próximos a aeroportos, otimizando a alocação de recursos.
- **Análises Avançadas e Capacidade Preditiva:** A granularidade e a organização dos dados nas tabelas de fatos e dimensões preparam o terreno para análises mais complexas e preditivas. A possibilidade de construir a matriz de transição para a Cadeia de Markov a partir da `FatoLocacoes` é um diferencial, permitindo a previsão de ocupação de pátios e, consequentemente, a otimização logística, a realocação proativa de veículos e a minimização de custos com ociosidade ou falta de frota.
- **Escalabilidade e Manutenibilidade:** O design com chaves surrogadas e o tratamento para Slowly Changing Dimensions na `DimCliente` demonstram uma arquitetura pensada para a evolução e manutenção do DW a longo prazo. Isso garante que a solução possa absorver o crescimento das operações e lidar com mudanças nos dados transacionais sem comprometer a integridade histórica e a performance analítica.

Em suma, o sistema projetado trará diversos benefícios para a operação diária das locadoras, desde a otimização da gestão de ativos (sabendo a localização e status dos veículos) até o planejamento estratégico (previsão de demanda). A implementação física e a monitorização periódica da performance e integridade dos dados serão os próximos passos cruciais para garantir o sucesso contínuo desta solução de Business Intelligence e maximizar o retorno sobre o investimento para o consórcio.