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 Final: Construção do Data Warehouse e Processo ETL

1. Introdução

Este relatório detalha a arquitetura e as funcionalidades da solução de Data Warehouse (DW) projetada para a associação de seis empresas independentes de locação de veículos. O objetivo central do projeto foi criar uma plataforma de dados unificada para superar a heterogeneidade dos sistemas operacionais (OLTP) existentes, permitindo a geração de relatórios gerenciais globais e análises estratégicas que antes eram inviáveis.

2. Comentários sobre o Desenvolvimento (Escolhas e Observações)

Esta seção detalha as decisões de projeto, os desafios encontrados e as conclusões sobre o processo de desenvolvimento da solução de Data Warehouse, atendendo ao requisito de um comentário geral sobre o trabalho.

Escolhas de Arquitetura e Modelagem

A escolha central do projeto foi a implementação de uma arquitetura de três camadas (Staging, Data Warehouse, Relatórios), que é um padrão robusto e escalável. Para o modelo do DW, optou-se pelo **Esquema Estrela (Star Schema)** devido à sua simplicidade, performance otimizada para ferramentas de BI e fácil compreensão pelos usuários de negócio.

A decisão de criar três tabelas de fatos distintas (`fato_locacoes`, `fato_reservas`, `fato_ocupacao_patio`) foi tomada para separar claramente os diferentes processos de negócio (locações, demanda e ocupação), tornando as análises mais diretas e o modelo mais manutenível. Para a `fato_ocupacao_patio`, a escolha de um modelo transacional (registrando

entradas e saídas) em vez de um snapshot foi estratégica, pois oferece maior flexibilidade para analisar o fluxo e a ocupação em qualquer período de tempo.

Problemas e Desafios Encontrados: O maior desafio, como esperado em qualquer projeto de integração, foi a **heterogeneidade das fontes de dados**. Durante a análise, ficou claro que não havia um padrão entre os sistemas, resultando em problemas como:

1. **Unificação de Entidades:** Um cliente no sistema da Empresa 1 poderia ter um ID diferente no sistema da Empresa 2. A solução foi definir uma chave de negócio (o `cpf_cnpj`) para unificar esses registros na `dim_cliente`.
2. **Conformação de Atributos:** Um mesmo conceito de negócio era representado de formas distintas. O exemplo clássico foi o campo de mecanização do veículo, que variava entre `VARCHAR` ('Manual', 'Automatica') e `BOOLEAN`. Isso foi resolvido na camada de transformação (`transformacao.sql`), que aplica uma regra `CASE` para padronizar todos os valores em um formato único ('Manual' ou 'Automático') na `dim_veiculo`.
3. **Extração de Dados Complexos:** Em alguns modelos, obter um dado simples exigia `JOINs` complexos. Por exemplo, no sistema do Grupo 4, para obter o nome e o CPF de um cliente, foi necessário unir três tabelas (`cliente`, `pessoa_fisica`, `pessoa_juridica`). A camada de extração foi projetada para lidar com essa complexidade, simplificando o modelo para as fases seguintes.

Observações Gerais e Conclusão do Processo: O processo ETL foi desenhado para ser idempotente, ou seja, pode ser reexecutado sem gerar duplicatas, graças à limpeza da Staging Area a cada ciclo e à lógica de verificação nas fases de transformação e carga. A criação de uma `procedure` para popular a `dim_tempo` e o uso de `TRANSACTION` na carga dos fatos são exemplos de práticas adotadas para garantir a robustez e a integridade do DW.

Durante o desenvolvimento deste projeto o grupo percebeu que o valor de um Data Warehouse não está apenas no armazenamento de dados, mas no processo criterioso de ETL que os transforma. Superar os desafios de integração foi fundamental para converter dados operacionais, fragmentados e inconsistentes em um ativo de informação centralizado e confiável, que habilita a empresa a responder perguntas de negócio complexas e a tomar decisões estratégicas com segurança.

3. Análise das Fontes de Dados e Justificativa do ETL

Conforme solicitado, a construção do processo ETL foi baseada na integração de quatro esquemas de banco de dados OLTP distintos, que representam os sistemas transacionais de cada empresa. A heterogeneidade entre esses sistemas foi o

principal motivador para as decisões de design da Staging Area e da camada de Transformação.

Fontes de Dados Utilizadas:

- **Fonte 1 (Nosso Grupo):** `modelo-fisico.sql`
- **Fonte 2 (Grupo Manhães):** `Manhaes.sql`
- **Fonte 3 (Grupo Medeiro):** `Medeiro.sql`
- **Fonte 4 (Grupo Kauer):** `Kauer.sql`

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.

4. Descrição do Modelo Dimensional (Star Schema)

O Data Warehouse foi projetado utilizando um **Esquema Estrela**, que consiste em tabelas de fatos (armazenando métricas) cercadas por tabelas de dimensão (armazenando contexto).

4.1. Dimensões

- **`dim_cliente`**
    - **Propósito:** Criar um cadastro mestre e único de todos os clientes das quatro empresas, permitindo uma visão 360º.
    - **Justificativa dos Campos:**
        - `id_cliente (PK)`: Chave substituta numérica para garantir performance e independência das chaves originais.

        - `cpf_cnpj`: Chave natural de negócio, usada para identificar e unificar clientes de diferentes fontes.
        - `cidade`, `estado`: Permitem a análise de locações e reservas por origem geográfica do cliente.
    - **Mapeamento das Fontes:** Os dados são extraídos das tabelas `CLIENTE`, `clientes`, e da junção de `cliente` com `pessoa_fisica`/`pessoa_juridica` das fontes OLTP.
- **`dim_veiculo`**
    - **Propósito:** Consolidar toda a frota de veículos em um catálogo único.
    - **Justificativa dos Campos:**
        - `id_veiculo (PK)`: Chave substituta.
        - `placa`: Chave natural de negócio para identificação única do veículo.
        - `grupo`, `tipo_mecanizacao`: Campos conformados, padronizados durante o ETL para permitir agrupamentos consistentes em relatórios.
    - **Mapeamento das Fontes:** Os dados são extraídos das tabelas `VEICULO` e `veiculos`, com `JOINs` nas tabelas de grupo de cada fonte.
- **`dim_patio`**
    - **Propósito:** Manter uma lista mestra e consistente dos seis pátios compartilhados.
    - **Justificativa dos Campos:**
        - `id_patio (PK)`: Chave substituta.
        - `nome`: Nome padronizado do pátio (ex: "Aeroporto do Galeão"), usado para garantir que todas as análises se refiram ao mesmo local de forma consistente.
    - **Mapeamento das Fontes:** Os dados são extraídos das tabelas `PATIO` e `patios` das fontes.
- **`dim_tempo`**
    - **Propósito:** Permitir a navegação e análise dos fatos ao longo do tempo em diferentes granularidades.
    - **Justificativa dos Campos:** Contém atributos como `ano`, `mes`, `dia`, `trimestre`, etc., para fatiar os dados sem a necessidade de cálculos de data complexos em tempo de consulta.
    - **Mapeamento das Fontes:** Esta dimensão não vem das fontes, mas é gerada e pré-populada por uma `procedure` no script de transformação.

4.2. Tabelas de Fatos

- **`fato_locacoes`**
  - **Propósito:** Registrar cada evento de locação concluída ou em andamento. É o fato principal do negócio.
  - **Justificativa dos Campos:**
    - `id_locacao (PK)`: Chave primária do fato.
    - `id_cliente`, `id_veiculo`, `id_patio_retirada`, `id_patio_devolucao`, `id_tempo (FKs)`: Chaves estrangeiras que conectam o fato às dimensões, permitindo a análise contextual.
    - `tempo_locacao`, `valor_final`: Métricas quantitativas e aditivas que medem a performance do negócio.
  - **Mapeamento das Fontes:** Os dados são extraídos das tabelas `LOCACAO`, `contrato` e `locacoes`, com `JOINs` nas tabelas de cobrança para obter os valores finais.
- **`fato_reservas`**
  - **Propósito:** Registrar cada evento de reserva, permitindo a análise da demanda futura.
  - **Justificativa dos Campos:**
    - `id_reserva (PK)`: Chave primária do fato.
    - `id_cliente`, `id_patio`, `id_tempo (FKs)`: Conectam o fato às dimensões.
    - `tempo_antecedencia`, `tempo_duracao_previsto`: Métricas que medem o comportamento do cliente ao reservar.
  - **Mapeamento das Fontes:** Os dados são extraídos das tabelas `RESERVA` e `reservas`.
- **`fato_ocupacao_patio`**
  - **Propósito:** Registrar as movimentações de entrada e saída de veículos nos pátios, permitindo a análise de ocupação.
  - **Justificativa dos Campos:**
    - `id_patio`, `id_tempo (FKs)`: Conectam o fato às dimensões.
    - `grupo`, `origem_empresa`: Dimensões degeneradas que permitem analisar a ocupação por tipo de carro e se pertence à frota própria ou de associados.
    - `qtd_veiculos`: Métrica transacional (`+1` para entrada, `-1` para saída) que, quando agregada ao longo do tempo, resulta no saldo de ocupação.
  - **Mapeamento das Fontes:** Os dados são derivados da tabela `stg_locacoes`, registrando um evento para a `data_retirada_real` (saída) e outro para a `data_devolucao_real` (entrada).

5. Cobertura dos Requisitos de Negócio e Conclusão

A solução projetada atende diretamente a todos os requisitos de relatórios e análises. Os scripts (`extracao.sql`, `transformacao.sql`, `carga.sql`, `relatorios.sql`) formam um pipeline completo e funcional que transforma dados operacionais brutos em informações estratégicas e acionáveis.

O modelo dimensional em estrela, juntamente com o processo ETL, permite superar os desafios da heterogeneidade das fontes de dados. 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 da associação.