**3.1 – Modelagem e Criação do Data Warehouse**

A primeira etapa do projeto consistiu na **modelagem e preparação do ambiente do Data Warehouse (DW)**, com o objetivo de consolidar e estruturar os dados de produção offshore de maneira confiável para futuras análises de qualidade de dados.

**Definição do Modelo**

Optou-se pela utilização de um **esquema estrela**, por ser uma abordagem amplamente utilizada em ambientes de Business Intelligence devido à sua **simplicidade, desempenho em consultas analíticas** e **facilidade de entendimento por usuários de negócio**.

O modelo foi estruturado com:

* **Uma Tabela Fato (fato\_projetos)**, contendo as informações centrais dos projetos de produção offshore (região, país, operador, profundidade, status, entre outros).
* **Duas Tabelas Dimensão (dim\_dutos e dim\_subsea\_trees)**, representando, respectivamente, os dutos e as árvores submarinas associadas a cada projeto.

Os relacionamentos foram definidos como **1 : N** (um projeto pode conter vários dutos e várias árvores submarinas), conforme ilustrado no diagrama abaixo:

**Diagrama do Esquema Estrela**

*(Inserir aqui o diagrama que geramos no Dia 1)*

**Criação Física no Banco de Dados**

A modelagem foi implementada no **PostgreSQL**, com a criação do banco de dados **dw\_qualidade\_og** e das respectivas tabelas.  
Os principais campos foram definidos com tipos de dados apropriados (inteiros para chaves e medidas, texto para atributos descritivos e números decimais para profundidade e distâncias).

As tabelas foram criadas respeitando os relacionamentos por meio de **chaves primárias e estrangeiras**, garantindo integridade referencial desde o início.

Um exemplo resumido das tabelas é apresentado abaixo:

| **Tabela** | **Principais Campos** |
| --- | --- |
| **fato\_projetos** | id\_projeto (PK), nome\_projeto, regiao, pais, bloco, campo, operador, tipo, status, profundidade\_m, latitude, longitude, distancia\_costa\_km |
| **dim\_dutos** | id\_duto (PK), id\_projeto (FK), nome\_duto, funcao, tipo\_duto, status, diametro\_polegada, comprimento\_km, ano\_instalacao |
| **dim\_subsea\_trees** | id\_tree (PK), id\_projeto (FK), status, ano\_instalacao |

**3.2 – Etapa de Extração, Transformação e Carga (ETL)**

Após a modelagem do Data Warehouse e a criação das tabelas no PostgreSQL, foi realizada a etapa de **Extração, Transformação e Carga (ETL)** utilizando a ferramenta **Pentaho Data Integration (PDI)**, também conhecida como Spoon.

O objetivo desta etapa foi **popular as tabelas do DW com dados provenientes dos três arquivos Excel fornecidos**:

* **Projetos.xlsx** (dados principais dos projetos);
* **Dutos.xlsx** (dados dos pipelines associados aos projetos);
* **Subsea Trees.xlsx** (dados das árvores submarinas).

**Processo Executado**

1. **Conexão com o Banco de Dados**  
   Inicialmente, foi criada uma conexão nativa no Pentaho com o banco de dados PostgreSQL **dw\_qualidade\_og**, permitindo a gravação direta dos dados tratados.
2. **Carga da Tabela Fato (fato\_projetos)**
   * Foi criado um fluxo simples, conectando o **step “Microsoft Excel Input”** ao **“Table Output”**.
   * Foram mapeados apenas os campos correspondentes à modelagem do DW, descartando colunas desnecessárias.
   * Foram carregados **585 registros**, correspondentes aos projetos de produção offshore.
3. **Carga da Dimensão Dutos (dim\_dutos)**
   * Para garantir a integridade referencial, foi utilizado o **step “Database Lookup”**, que buscou automaticamente o **id\_projeto** na tabela fato\_projetos, com base no nome do projeto presente no Excel.
   * Após o mapeamento dos campos necessários (nome\_duto, funcao, tipo\_duto, status, diametro\_polegada, comprimento\_km, ano\_instalacao), foram inseridos **2267 registros**.
4. **Carga da Dimensão Subsea Trees (dim\_subsea\_trees)**
   * Processo semelhante ao dos dutos, com o uso do **Database Lookup** para vincular cada árvore submarina ao projeto correto.
   * Foram mapeados apenas três campos: id\_projeto, status e ano\_instalacao.
   * Foram carregados **2651 registros**.

**Resultados da Etapa**

Ao final do processo, o Data Warehouse ficou completamente populado, respeitando a integridade referencial entre as tabelas. Os resultados podem ser resumidos da seguinte forma:

| **Tabela** | **Quantidade de Registros** |
| --- | --- |
| **fato\_projetos** | 585 |
| **dim\_dutos** | 2267 |
| **dim\_subsea\_trees** | 2651 |

3.3 **Análises de Qualidade de Dados (Data Quality - DQ)**

Com o objetivo de avaliar a confiabilidade e completude dos dados do *data warehouse* de projetos offshore, foram desenvolvidas e executadas quatro análises de qualidade de dados (**DQ01 a DQ04**). Cada uma aborda um aspecto específico da qualidade, permitindo identificar lacunas, inconsistências e oportunidades de melhoria na base.

**DQ01 – Completude dos Dados**

A análise de completude buscou medir o percentual de preenchimento de cada campo crítico da tabela de projetos (fato\_projetos).  
Para isso, cada campo foi avaliado individualmente, atribuindo **1** para valores preenchidos e **0** para valores nulos ou em branco. Em seguida, os valores foram agregados e transformados em percentual sobre o total de registros.

Os campos avaliados incluíram:

* nome\_projeto, regiao, pais, bloco, campo, operador, tipo, status, profundidade, latitude, longitude, distancia\_costa\_km.

O resultado final foi gravado na tabela dq\_completude\_fato, possibilitando análises posteriores. Como exemplo, obteve-se que o campo **bloco** apresentou apenas **46,32%** de preenchimento, enquanto campos como **tipo** e **status** atingiram **100%**.

**DQ02 – Consistência dos Dados**

A análise de consistência teve como objetivo validar se os valores preenchidos respeitavam regras lógicas pré-definidas. Foram criadas regras simples, mas eficazes, para cada campo crítico numérico:

* **Profundidade (profundidade)**: valores não nulos e maiores que 0;
* **Latitude (latitude)**: valores entre **-90** e **90**;
* **Longitude (longitude)**: valores entre **-180** e **180**;
* **Distância da Costa (distancia\_costa\_km)**: valores não nulos e maiores ou iguais a 0.

Cada regra também gerou valores binários (1 = consistente, 0 = inconsistente), e os percentuais foram calculados de forma similar ao DQ01.  
Por exemplo, identificou-se que apenas **64,27%** dos registros possuem **latitude** e **longitude consistentes**, o que evidencia oportunidades de melhoria na padronização e controle dessas informações.

Os resultados foram armazenados na tabela dq\_consistencia\_fato.

**DQ03 – Duplicidade de Registros**

Para identificar possíveis duplicidades, foi criada uma **chave composta** concatenando os campos **nome\_projeto + pais**. Essa combinação foi considerada única para cada projeto.

Em seguida, o número de ocorrências de cada chave foi contado. Registros com contagem **superior a 1** foram classificados como duplicados.

Os resultados foram armazenados na tabela dq\_duplicidade\_fato, indicando tanto **quais são os projetos duplicados** quanto **quantas vezes cada um aparece**.  
Por exemplo:

* "Enchova\_Brazil" → 2 registros
* "Tartaruga\_Brazil" → 2 registros

**DQ04 – Classificação dos Projetos por Profundidade**

Por fim, foi realizada uma classificação dos projetos em três categorias, com base na profundidade de operação:

* **Shallow Water** → profundidade < 200 metros;
* **Deepwater** → profundidade ≥ 200 metros;
* **Indefinido** → profundidade nula ou inválida.

Os registros foram agregados por categoria e armazenados na tabela dq\_classificacao\_fato.  
O resultado geral foi:

* **Shallow Water → 322 projetos**
* **Deepwater → 251 projetos**
* **Indefinido → 12 projetos**

Essa classificação possibilita análises mais estratégicas sobre a distribuição dos projetos e também auxilia no monitoramento de possíveis lacunas de preenchimento (categoria "Indefinido").

**Conclusão Parcial**

Essas quatro análises, em conjunto, fornecem uma visão abrangente da qualidade dos dados, permitindo:

* Identificar campos mais críticos em termos de preenchimento;
* Avaliar se os valores respeitam regras lógicas básicas;
* Detectar duplicidades que podem distorcer análises operacionais;
* Classificar projetos de forma estratégica, destacando lacunas.

A execução periódica desses indicadores é recomendada para garantir a evolução da qualidade da base.