Data Warehouse e Visualização para Análise de Qualidade em Dados de Projetos Upstream

Matheus Müller Barbosa

PUC – Rio. Pontifícia Universidade Católica do Rio de Janeiro

Curso de Pós-Graduação \*Business Intelligence Master\*

**Abstract.** This work aims to develop an analytical environment focused on assessing data quality in offshore oil and gas projects in South America. The project involves the construction of a Data Warehouse (DW) using data extracted from spreadsheets related to projects, pipelines, and subsea trees, structured in a star schema and loaded into a PostgreSQL database. ETL processes were developed using Pentaho to apply rules that evaluate the main pillars of data quality: completeness, consistency, duplication, and classification. Finally, the processed data was visualized in a Power BI dashboard, providing a practical tool for diagnostics and decision-making. The results demonstrate the solution’s effectiveness in identifying data issues and supporting data governance in the oil and gas industry.

**Resumo.** Este trabalho tem como objetivo desenvolver um ambiente analítico voltado à avaliação da qualidade dos dados utilizados em projetos offshore de exploração e produção de petróleo na América do Sul. A proposta consiste na construção de um Data Warehouse (DW) com base em dados extraídos de planilhas relacionadas a projetos, dutos e árvores submarinas, modelados em uma estrutura estrela e carregados em um banco de dados PostgreSQL. Por meio de transformações ETL desenvolvidas no Pentaho, foram aplicadas regras para analisar os principais pilares da qualidade da informação: completude, consistência, duplicidade e classificação. Por fim, as informações tratadas foram visualizadas em um dashboard no Power BI, oferecendo uma ferramenta prática para diagnóstico e tomada de decisão. Os resultados demonstram a eficácia da solução na identificação de falhas e na promoção da governança de dados no setor de óleo e gás.

# 1. Introdução

# Com o avanço do uso de dados como ativo estratégico nas organizações, a preocupação com a qualidade da informação torna-se cada vez mais crítica, especialmente em setores complexos como o de óleo e gás. Projetos offshore envolvem grandes investimentos, alto risco e requisitos rigorosos de planejamento e execução. Nesse contexto, inconsistências, duplicações ou ausência de dados podem gerar prejuízos operacionais e financeiros. Para mitigar esses riscos, este trabalho propõe a construção de um Data Warehouse (DW) voltado à análise da qualidade dos dados em projetos de desenvolvimento offshore na América do Sul, com foco nos atributos de completude, consistência, duplicidade e classificação.

# 2. Objetivo do Trabalho

O objetivo principal deste trabalho é criar um ambiente analítico para monitoramento e diagnóstico da qualidade dos dados relacionados a projetos de exploração e produção offshore. Para isso, foram utilizadas ferramentas como Pentaho, PostgreSQL e Power BI para realizar a ingestão, tratamento, modelagem e visualização dos dados. O trabalho também visa demonstrar, por meio de um dashboard interativo, os principais problemas de qualidade dos dados, fornecendo subsídios para ações corretivas e melhoria contínua na governança da informação.

# 3. Metodologia

A metodologia adotada neste trabalho final compreendeu a criação de um ambiente de Business Intelligence com foco na avaliação da qualidade dos dados provenientes de projetos offshore. O processo foi dividido em cinco etapas principais: coleta e entendimento dos dados, modelagem dimensional, processo ETL, análise da qualidade dos dados e visualização por meio de dashboard. Cada etapa foi cuidadosamente planejada e executada com o uso de ferramentas robustas e aderentes ao mercado, como PostgreSQL, Pentaho Data Integration (PDI) e Power BI.

**3.1. Coleta e compreensão dos dados**

A base de dados utilizada foi composta por três planilhas fornecidas em formato Excel: uma contendo informações sobre projetos offshore (nome, localização, status, profundidade etc.), outra com dados sobre dutos submarinos (diâmetro, comprimento, tipo, localização e projeto associado), e uma terceira com dados sobre árvores submarinas (quantidade, tipo, localização, projeto associado, entre outros).

Foi realizada uma análise exploratória para compreender o significado de cada coluna e identificar os principais atributos que seriam utilizados na modelagem dimensional e nas regras de qualidade.

**3.2. Modelagem Dimensional**

Foi adotado o modelo estrela para a estruturação do Data Warehouse, com a tabela fato concentrando os projetos offshore (fato\_projetos) e três tabelas dimensionais: dim\_dutos, dim\_subsea\_trees e a dim\_temporal para futura expansão. Essa escolha se deu pela sua simplicidade e eficiência em consultas analíticas.

A modelagem considerou chaves substitutas e a normalização de dados categóricos. Cada dimensão foi associada à fato\_projetos por meio do nome do projeto, utilizado como chave de junção no processo de ETL.

**3.3. Processo ETL (Extract, Transform, Load)**

As transformações ETL foram desenvolvidas utilizando o Pentaho Data Integration (Spoon). Cada etapa do processo foi cuidadosamente projetada para garantir a integridade e padronização dos dados.

• ETL01 - GRAVA-PROJETOS: Responsável por extrair os dados da planilha de projetos e carregá-los na tabela fato\_projetos do DW.

• ETL02 - GRAVA-DUTOS: Carrega os dados de dutos na dim\_dutos, relacionando-os com os projetos.

•ETL03 - GRAVA-TREES: Carrega os dados de árvores submarinas na dim\_subsea\_trees.

Todas as transformações aplicam conversões de tipos, validação de campos obrigatórios, remoção de linhas inválidas e junção com a tabela de projetos.

A screen shot of a computer

AI-generated content may be incorrect.

Figura 1) Imagem ilustrando carregamento da tabela ‘Trees’ como exemplo

**3.4. Análise da Qualidade dos Dados**

Foram criadas quatro transformações ETL adicionais voltadas exclusivamente à análise da qualidade dos dados da tabela fato:

* **DQ01-COMPLETUDE-FATO**: Verifica a porcentagem de preenchimento dos campos-chave para identificar ausências de dados.
* **DQ02-CONSISTENCIA-FATO**: Aplica regras de negócio simples (ex.: valores negativos em profundidade, latitude/longitude inválidas) para verificar consistência.
* **DQ03-DUPLICIDADE-FATO**: Identifica registros duplicados a partir da combinação de nome do projeto e país.
* **DQ04-CLASSIFICACAO-FATO**: Classifica os projetos em “Shallow Water”, “Deepwater” ou “Indefinido” com base na profundidade (<200m ou >=200m).

Essas transformações gravam os resultados em tabelas auxiliares no DW, facilitando sua posterior visualização e auditoria.

A diagram of a diagram

AI-generated content may be incorrect.

Figura 2) Ilustração da Transformação DQ01-COMPLETUDE-FATO

**3.5. Automatização com Job Geral**

Para garantir a execução sequencial de todas as transformações, foi criado um job chamado DQ\_JOB\_GERAL, que executa:

1. ETL01-GRAVA-PROJETOS

2. ETL02-GRAVA-DUTOS

3. ETL03-GRAVA-TREES

4. DQ01-COMPLETUDE-FATO

5. DQ02-CONSISTENCIA-FATO

6. DQ03-DUPLICIDADE-FATO

7. DQ04-CLASSIFICACAO-FATO

Essa execução integrada permite a atualização automática do Data Warehouse com novos dados. A figura 3 ilustra o sequenciamento do Job.

A computer screen shot of a line

AI-generated content may be incorrect.

Figura 3) Ilustração de Funcionamento do Job DQ\_JOB\_GERAL

**3.6. Visualização em Power BI**

As tabelas do banco PostgreSQL foram conectadas ao Power BI Desktop, onde foi criado um dashboard interativo contendo os seguintes indicadores:

• Percentual de completude por principais campos (latidude, longitude, distancia e profundidade).

• Percentual de consistência por regra.

• Quantidade de registros duplicados e sua identificação.

• Distribuição de projetos entre Shallow Water e Deepwater.

• Mapa gerado através das informações de longitude e latitude.

A visualização permitiu a análise rápida da qualidade dos dados, facilitando a tomada de decisões sobre as necessidades de correção e aprimoramento dos processos de coleta de informações no setor offshore.

A screenshot of a computer screen

AI-generated content may be incorrect.

Figura 4) Ilustração do Dashboard de Data Quality

**4. Resultados**

A execução do projeto resultou na construção bem-sucedida de um Data Warehouse voltado à análise da qualidade de dados em projetos offshore de óleo e gás na América do Sul. O modelo dimensional implementado permitiu consolidar e relacionar dados de três fontes principais: projetos, dutos e árvores submarinas.

Após o carregamento das dimensões e da tabela fato, foi realizado o processo de validação da qualidade dos dados por meio de quatro análises distintas:

a. Completude: Avaliou a proporção de campos preenchidos em relação ao total esperado, considerando atributos-chave como nomes de projeto, datas e localização. O resultado geral de completude se mostrou satisfatório, com mais de 90% de preenchimento na maioria dos campos críticos.

b. Consistência: Verificou a coerência entre campos dependentes, como a relação entre datas de início e fim dos projetos. A aplicação de regras de negócio permitiu identificar inconsistências temporais e falhas lógicas, que foram devidamente sinalizadas.

c. Duplicidade: Identificou registros possivelmente duplicados na tabela fato com base em combinações de atributos-chave. A análise revelou seis duplicidades, que foram destacadas para posterior verificação e tratativa.

d. Classificação: Atribuiu uma classificação aos projetos conforme sua profundidade operacional: “Shallow Water” para profundidades abaixo de 200 metros, “Deepwater” para profundidades iguais ou superiores a esse valor, e “Indefinido” para registros sem informação. Esta classificação foi útil tanto para a análise dos dados quanto para a visualização dos resultados.

O dashboard desenvolvido no Power BI consolidou essas informações em gráficos e indicadores visuais que facilitam a compreensão da qualidade dos dados. Foram criados cards com índices de completude, consistência e duplicidade, além de gráficos com a distribuição de projetos por classificação e por país. Essa visualização permitiu identificar pontos críticos e direcionar ações corretivas.

A execução automática do processo ETL via job no Pentaho garantiu a reprodutibilidade da análise, permitindo atualizações futuras com facilidade. O fluxo completo de extração, transformação, carga e validação dos dados foi integrado com sucesso.

**5. Conclusão**

Este trabalho demonstrou a importância e a viabilidade da implementação de um Data Warehouse para fins de monitoramento da qualidade de dados em projetos offshore. Por meio da aplicação de técnicas de ETL, modelagem dimensional e visualização interativa, foi possível consolidar informações dispersas e gerar indicadores valiosos para o processo decisório.

As análises de qualidade permitiram identificar falhas, incoerências e lacunas nos dados, servindo como base para melhorias no processo de coleta e tratamento de informações. Além disso, a automação do processo assegura que as futuras atualizações mantenham o mesmo padrão de rigor analítico.

**6. Referências**

LAGO, Karine e ALVES Laennder. Dominando o Power BI. 3. Ed. São Paulo: DATAB Inteligência e Estratégia, 2020.

<https://www.postgresql.org/docs/>

<https://help.hitachivantara.com/Documentation/Pentaho/9.2>

Conhecimentos adquiridos na abordagem da disciplina BI - Business Intelligence do curso de pós-graduação BI MASTER da PUC – RIO.