AGTU – THE GLOBAL UNIVERSITY

ESTUDO DE CASO: OTIMIZAÇÃO DE PROCESSOS DE ETL EM UMA EMPRESA DE E-COMMERCE

AUTOR: MARCIO LUCENA MAGALHAES

Introdução

A SportShop, uma empresa de grande porte de e-commerce, focada em venda online de produtos esportivos para consumidores que buscam a prática de esportes, atividades físicas e qualidade de vida. Com milhões de usuários e milhares de transações diárias, a empresa enfrenta desafios significativos na gestão de grandes volumes de dados provenientes de diversas fontes, como planilhas, arquivos csv, sistemas de vendas, CRM, plataformas de marketing, entre outros. Entre os sistemas e plataformas utilizados podemos citar: Oracle para banco de dados, Amazon para armazenamento na nuvem, Plataforma VTEX de comércio para E-commerce, Salesforce para a gestão de relacionamento com o cliente, Sistema ERP e Gestão de Estoque da SAP, PagSeguro como sistema de pagamentos, as plataformas da Google e Meta para o marketing digital. Com o grande crescimento dos dados, afetando o desempenho e a eficiência dos seus processos de ETL (Extração, Transformação e Carga) tornou-se evidente melhorar a eficiência e a qualidade dos dados, visando obter insights mais rápidos e precisos para a tomada de decisões.

O objetivo deste estudo de caso é analisar o processo atual de ETL, identificar os principais pontos de ineficiência, e propor uma solução para otimizar esses processos. A metodologia utilizada fará uma análise do processo atual como

mapeamento do fluxo de ETL, medindo o desempenho e identificando gargalos. Terá uma proposta de melhoria como sugerir nova arquitetura ou modificar a existente, considerando novas tecnologias e escalabilidade. E a validação da proposta com a simulação de nova solução, comparando o desempenho e apresentando uma análise de custo-benefício.

Desenvolvimento

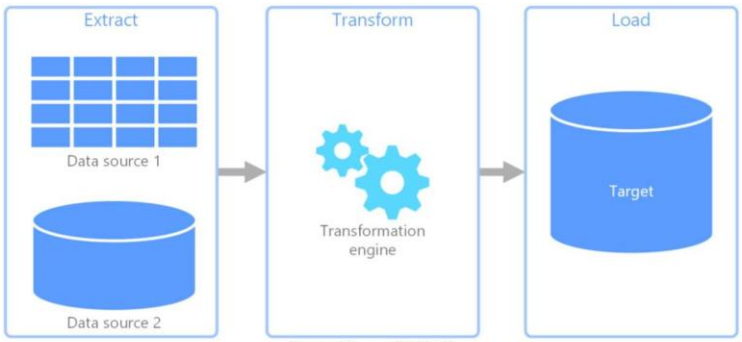
A SportShop tem um grande volume de transações diárias, milhares de produtos, milhões de clientes e interações no site, seguem dados fictícios descritos na tabela abaixo.

Descrição	Quantidade
Número de transações diárias de vendas	15.000 transações/dia
Número de Produtos	10.000 produtos
Dados de clientes, histórico de compras	2.000.000 clientes
Interações no site, buscas, cliques e emails	50.000 interações/dia

A empresa percebeu que seus processos de ETL estão se tornando um gargalo, atrasando as análises e impactando negativamente a tomada de decisões baseada em dados. Com o aumento do tempo de processamento, fonte de dados diversas, inconsistência de dados, a demora e erros na tomada de decisão, tem ocasionado:

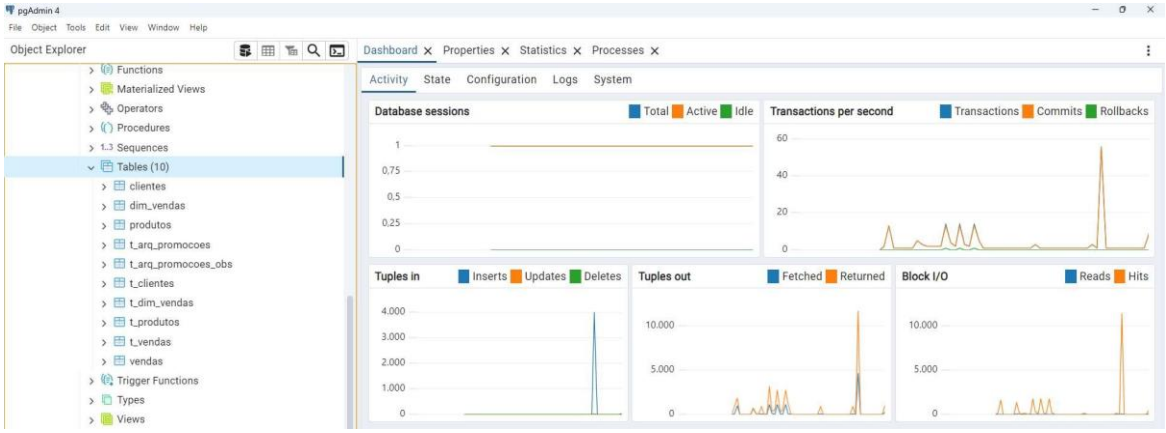
- Perdas de clientes num mercado altamente competitivo de e-commerce;
- Alto custo computacional devido ao tempo de processamento;
- Demora e erros na tomada de decisões.

O processo ETL extrai os dados da origem, transforma os dados de acordo com as regras de negócios e carrega os resultados numa base (Silva, 2019). É um processo que divide em três fases: Extração, Transformação e Carga. (Ferreira, Miranda, Machado, 2010)

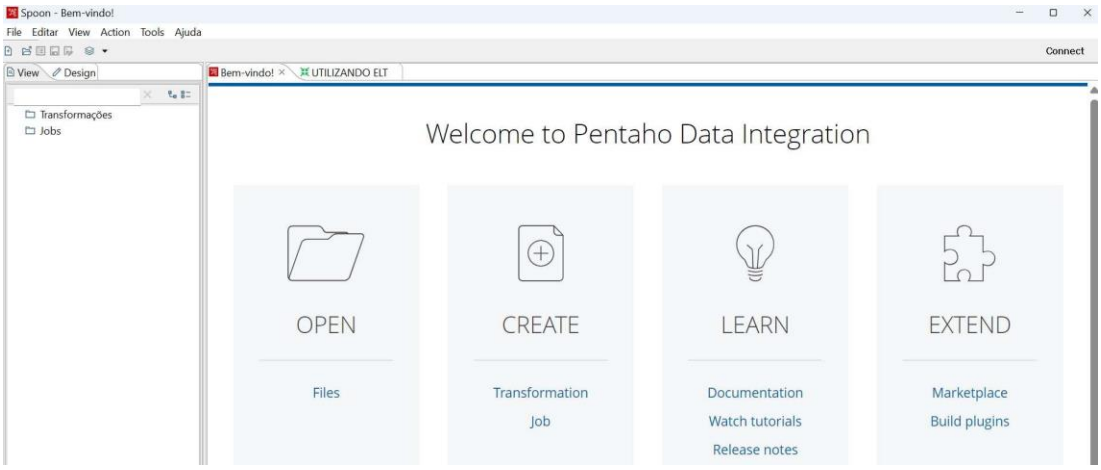


Fonte: Microsoft (2019)

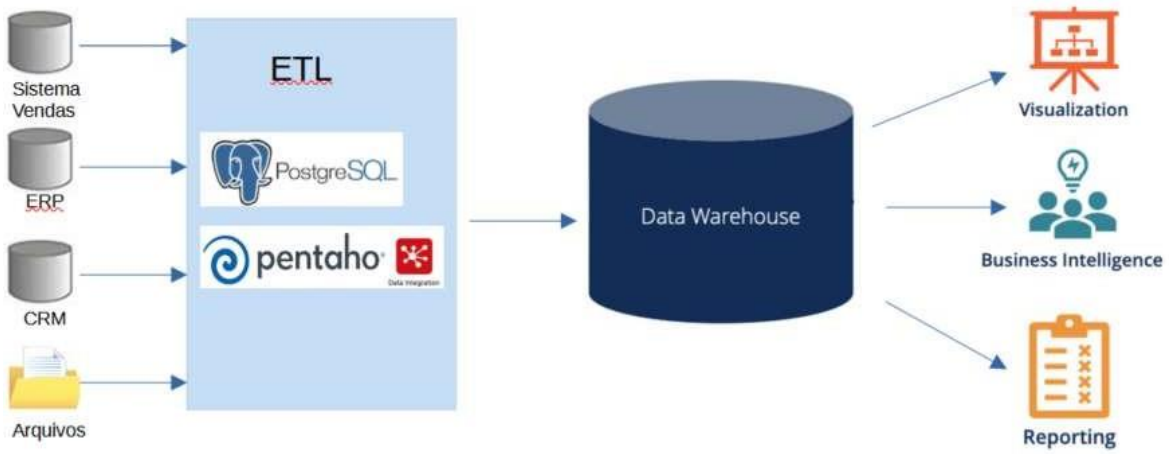
Para geração de dados sintéticos, foi utilizado a ferramenta Mockaroo. Para armazenar os dados, foi utilizado o Postgres por ser um SGBD open source e simples utilização, segue um print da ferramenta utilizada.



O Pentaho Data Integration foi utilizado para modelagem do ETL, pois é um software livre e possui todo seu código aberto (Mussa, Souza, Freire, 2018), permitindo uma integração completa, transformação de dados avançada, agendamento e orquestração de tarefas.



Segue o diagrama do processo atual de ETL: Extrair, Transformar e Carregar.



Segue uma simulação do fluxo atual de ETL utilizando o Pentaho.

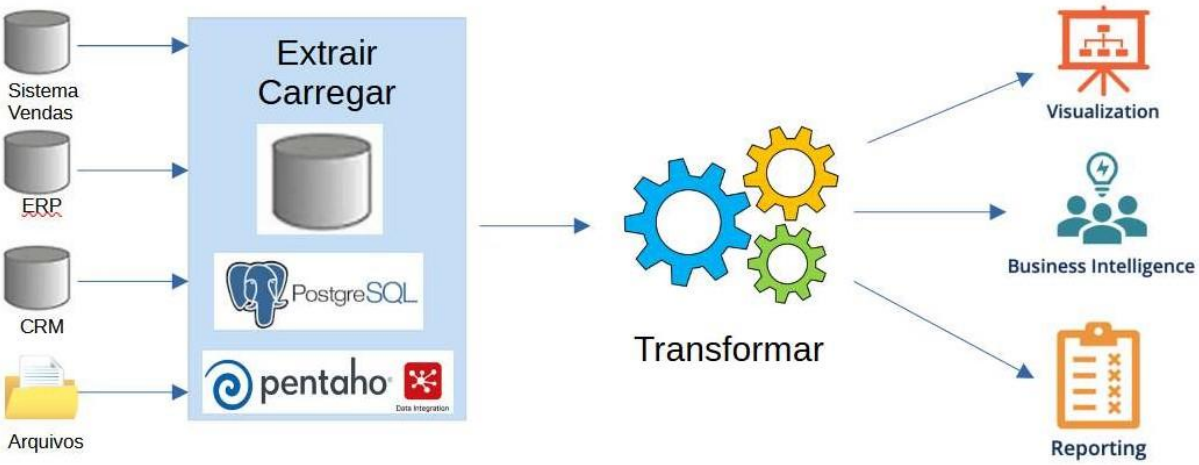


Como sugestão de melhoria para otimização do processo de ETL, foi utilizado a migração do ETL para o ELT (Extrair, Carregar, Transformar). Esta abordagem permite extrair e carregar dados, depois aplicar transformações sob demanda de acordo com a necessidade do negócio (Neto, 2022).

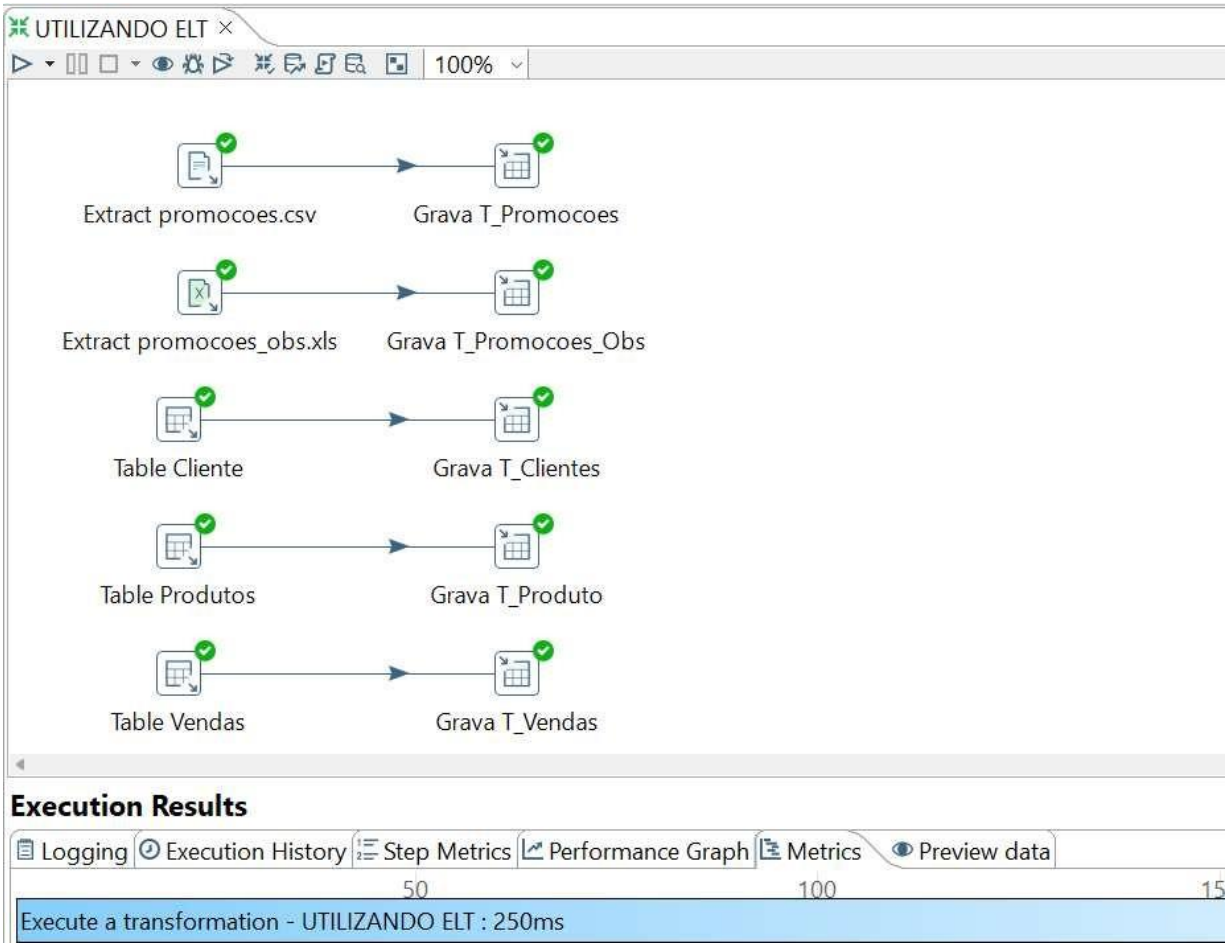
Numa comparação entre as abordagens ETL e ELT, o processo de ELT tem as características de ser mais rápido e com menor custo em relação ao ETL (Radovanovic, 2024).

Abordagem	ETL	ELT
Compatibilidade	Dados estruturados	Dados estruturados, semiestruturados e não estruturados
Velocidade	Mais lento	Mais rápido
Custo	Mais caro	Mais barato

A seguir, o diagrama do processo otimizado utilizando o ELT:



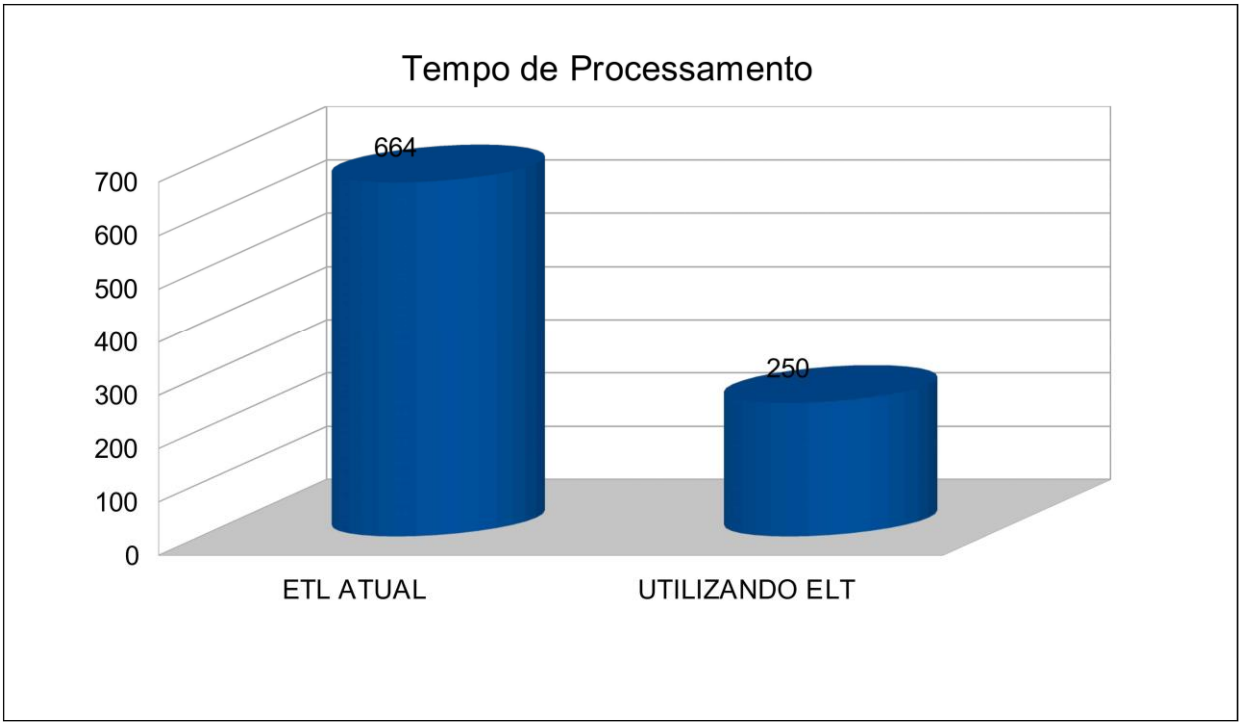
Segue uma simulação do fluxo atual de ELT utilizando o Pentaho, fazendo apenas as etapas de extração e carregamento.



Execution Results

Foi utilizado o tempo total de execução no processo como principal métrica como indicador de desempenho conforme tabela e gráfico abaixo. Tivemos um ganho expressivo de desempenho, pelo menos duas vezes melhor, o que significa que podem ser utilizados menos recursos computacionais com o novo processo utilizando o ELT.

PROCESSO	TEMPO DE PROCESSAMENTO	UNIDADE
ETL ATUAL	664	milisegundos
UTILIZANDO ELT	250	milisegundos
VELOCIDADE COM ELT	**265,60%**	**percentual**



Conclusão

A utilização do processo de ELT resultou em melhorias significativas em termos de eficiência operacional, menor custo, qualidade dos dados e disponibilidade rápida de informações para análise. O tempo de execução do processo como indicador de desempenho demonstrou que o que a abordagem ELT resultou num ganho de velocidade de 265,60%, pelo menos duas vezes maior que o ETL, e resultando também no menor custo de processamento. No trabalho desenvolvido foram instalados as ferramentas Postgres, Pentaho e LibreOffice.

Foram pesquisados artigos relacionados na internet e as referências foram citadas quando utilizadas no trabalho, contendo os links para os documentos.

As limitações encontradas foram o prazo curto e o tempo para aprendizado na utilização das ferramentas citadas.

REFERÊNCIAS

1. MUSSA, Munir de Sá; Souza, Saulo Chaves; Freire, Eduardo Francisco. Business Intelligence in Education: An Application of Pentaho Software, Revista de Produção e Desenvolvimento, Julho 2018. https://revistas.cefet-rj.br/index.php/producaoedesenvolvimento/article/view/ e274

2. FERREIRA, João; Miranda, Miguel; Machado, José. O processo ETL em sistemas Data Warehouse. Universidade do Minho, Portugal. 2010. https://hdl.handle.net/1822/11435

3. SILVA, André. Uso do processo de Extração, Transformação e Carga (ETL) voltado a visão de atributos qualificantes, associados ao auxílio à qualificação do

profissional de TI. Faculdade Doctum de Caratinga. 2019. https://dspace.doctum.edu.br/bitstream/123456789/3370/1/TCC2020.pdf

4. NETO, Genival. Um estudo sobre ferramentas de código aberto para carga de dados de documentos fiscais eletrônicos. Universidade Federal da Paraíba. 2022. https://repositorio.ufpb.br/jspui/bitstream/123456789/31582/1/Genival%20Jos%c3%a9%20de%20Moura%20Neto%20-%20TCC.pdf

5. RADOVANOVIC, Felipe. Desenvolvimento de pipelines para padronização, limpeza e enriquecimento de dados veiculares. Universidade Federal de Santa Catarina. 2024. https://repositorio.ufsc.br/bitstream/handle/123456789/254803/TCC_-_Felipe_Trindade_Radovanovic.pdf