V Encontro de Iniciação Científica e Tecnológica
V EnICT
ISSN: 2526-6772
IFSP – Câmpus Araraquara
22 e 23 de outubro de 2020

ERP de Pet Shop: técnica ETL na elaboração de um Data Warehouse de Vendas

ANA GUELFI ERBA[1], JOSÉ RODOLFO BELUZO[2], RENATA MARIA PORTO VANNI[3], TALES BOALIM[4]

[1] Discente do Curso Técnico em Informática Integrado ao Ensino Médio, Bolsa Fundação, IFSP Câmpus - Araraquara, ana.guelfi@aluno.ifsp.edu.br
[2,3] Docente do Instituto Federal de São Paulo – Câmpus Araraquara, [2]jrbeluzo@ifsp.edu.br, [3]rportovanni@ifsp.edu.br
[4] Gerente de Tecnologias e Inovação - Moura Informática

Área de conhecimento (Tabela CNPq): Banco de Dados – 1.03.03.03-0

RESUMO: Este trabalho refere-se à aplicação da técnica *Extract, Transform and Load* (ETL) em bases operacionais com dados transacionais de vendas de produtos e serviços de empresas de varejo, no ramo de Pet Shop. Foram criadas regras de extração, transformação e carregamento de dados na ferramenta de ETL *Pentaho Data Integration* visando a conversão dos dados brutos em dados estruturados, de modo a facilitar a geração de relatórios claros, objetivos e com dados padronizados. Assim, a estrutura das bases operacionais foi analisada para a elaboração de regras que integrem os dados das diversas bases do sistema de Pet Shop em um Data Warehouse (DW). Ao fim da integração, foram obtidas análises que auxiliam na tomada de decisões empresariais. Desse modo, são apresentadas algumas regras de transformação e padronização que constituem o processo de normalização do DW.

PALAVRAS-CHAVE: *Data Warehouse*; ETL; integração de dados; padronização de dados; regras de ETL.

INTRODUÇÃO

Segundo Elmasri e Navathe (2010), os dados são fatos que ao serem integrados e contextualizados, passam a agregar significado objetivo, pois geram informações consistentes que podem ser utilizadas principalmente em um cenário empresarial. Um *Data Warehouse* (DW) tem como principal função unir dados de diversas fontes de modo a integrá-los para facilitar as análises de dados e gerar relatórios com informações de um determinado grupo de dados que antes estava descentralizado, proporcionando um efetivo suporte à tomada de decisão em negócios (KIMBALL; CASERTA, 2004).

Na construção de um *Data Warehouse*, o ETL (do inglês, *Extract, Transform and Loading*) é um processo fundamental, pois garante a limpeza, completude, padronização e normalização para melhorar a qualidade dos dados que virão a ser integrados no DW. Dessa maneira, os dados são extraídos das fontes originais, transformados, por meio de regras, em um padrão que favoreça às futuras análises e, então, carregados no DW. Este processo engloba procedimentos de limpeza, integração e transformação de dados, sendo a etapa mais crítica e demorada na elaboração de um DW (FERREIRA et al., 2010).

Entre as diversas dificuldades para extração de dados de diferentes fontes, as especificidades intrínsecas de cada fonte dificultam a clareza dos resultados nas variadas consultas. A busca por padronização, limpeza e

integração de dados requer atenção a estas particularidades do sistema de origem, respeitando as relações entre tabelas e informações. Para que o DW atenda às demandas de negócio das empresas que o solicitam, é necessário descobrir os requisitos comerciais por meio de reuniões com a empresa, estudo das fontes de dados para predizer as futuras análises, verificar quais respostas as empresas pretendem obter com estas análises e, enfim, elaborar uma estrutura e regras de padronização que auxiliem a tomada de decisões (KIMBALL; ROSS, 2013).

Considerando os aspectos teóricos apresentados, o objetivo geral deste projeto de iniciação científica foi realizar a integração de dados de várias bases de um sistema de ERP para Pet Shops, de modo a propiciar efetivas análises de apoio a decisão, utilizando a ferramenta ETL chamada *Pentaho Data Integration* na construção de um *Data Warehouse*. Assim, analisou-se a estrutura e os dados do sistema para elaboração e aplicação de regras para a extração, transformação e carregamento dos dados das bases no DW.

FUNDAMENTAÇÃO TEÓRICA

Um *Data Warehouse* é comumente modelado como um esquema estrela ou um esquema floco de neve (INMON, 1995), ambos formados por tabelas de fatos e tabelas de dimensão. De acordo com Kimball e Caserta (2004), as tabelas de dimensão são responsáveis por armazenar dados que contextualizam os fatos, e a tabela de fatos armazena o principal objeto rentável de uma empresa, que na literatura é chamado de grão. Geralmente, nas empresas ERP, a venda é o grão.

O ambiente que une os dados operacionais ao *Data Warehouse* é chamado de ETL. Segundo Kimball e Ross (2013), ele é constituído por uma área de trabalho, conjuntos de dados e processos, dividido em três etapas principais, sendo elas a extração, a transformação e o carregamento, conforme a Figura 1. É essencial que o conjunto de regras ETL esteja de acordo com a estrutura desejada pela empresa que solicita o DW, além de respeitar relações entre as tabelas das fontes de dados (SEAH; SELAN, 2014).

A extração (*extract*), de acordo com Wijaya e Pudjoatmodjo (2015), é realizada após a definição do modelo do DW, extraindo e selecionando os dados da forma mais eficiente possível, pois é responsável por mapear os dados que constituem a tabela de fatos e as tabelas de dimensão. A etapa de transformação (*transform*) vem logo em seguida, realizando operações como a padronização do formato e tipo dos dados, o tratamento de dados inconsistentes e dos campos sem valores atribuídos, e a remoção de dados duplicados. Esta etapa pode ser chamada limpeza, porque garante a precisão e consistência dos dados e proporciona análises e gráficos claros e corretos. Finaliza-se o processo ETL com a etapa de carregamento (*load*), em que os dados já preparados e semi-integrados são carregados no *Data Warehouse*, respeitando sua estrutura.

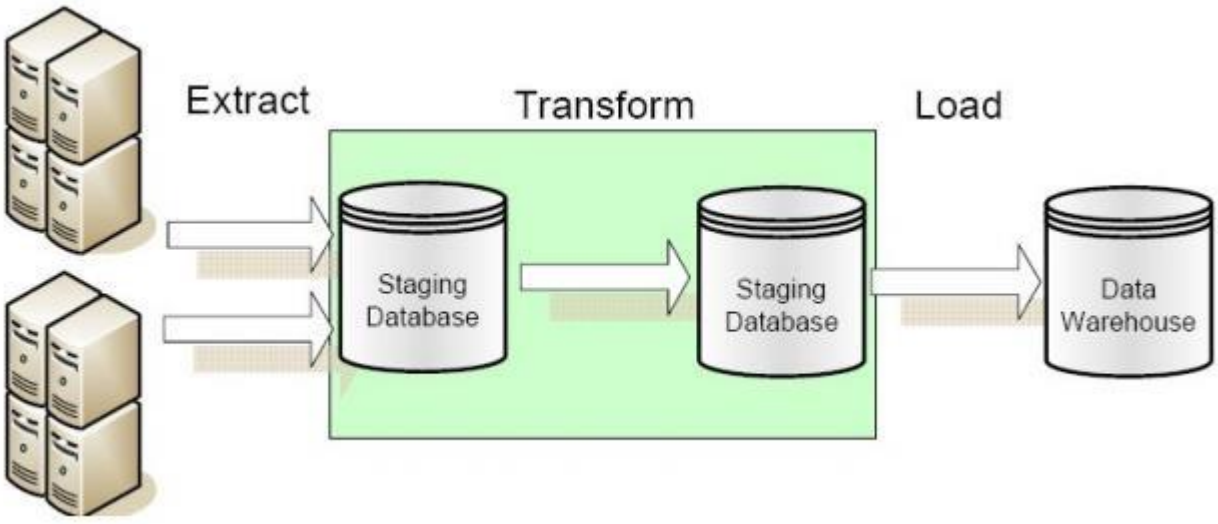

Figura 1: Processo de ETL para construção de um Data Warehouse.
Fonte: Ranjan (2009).

Segundo Hanlin, Xianzhen e Xianrong (2012), a etapa de transformação ocorre usando regras que geralmente possuem um conjunto de atributos de entrada que resultarão nos dados transformados e prontos para serem alocados no DW, podendo incluir mudança do tipo dos dados do campo, mapeamento de campos, alterar o nome dos campos originais para novos nomes no *Data Warehouse*, fazer operações com dados numéricos, unir campos através do merge, separar dados de tipo `string`, entre outras. Estes processos garantem dados padronizados e otimizados para a geração de relatórios e análises.

METODOLOGIA

A metodologia utilizada para o desenvolvimento do projeto foi constituída por leitura de livros e artigos sobre *Data Warehouse* e ETL, documentação dos modelos relacionais das bases de dados de Pet Shop, criação de um modelo de DW contendo as informações mais importantes relacionadas às vendas de uma empresa e que viriam a ser métricas interessantes para uma futura análise, escolha da ferramenta responsável pelas regras ETL e elaboração das regras, e avaliação dos resultados de forma qualitativa.

Dentre os processos metodológicos seguidos durante a realização do trabalho de pesquisa, destaca-se a elaboração das regras ETL. Estas regras foram definidas de modo que relacionassem a estrutura das bases de Pet Shop, as análises pretendidas e a forma como os dados originais se apresentavam. Após terem sido definidas, foram implementadas na ferramenta *Pentaho Data Integration*[1] desde a extração das fontes operacionais até o carregamento no *Data Warehouse*.

As bases operacionais apresentavam, muitas vezes, vários registros de um mesmo dado. No DW, para melhor normalização, é necessário remover estas repetições. Assim, uma regra elaborada para a remoção de valores duplicados, de acordo com a Figura 2. Para esta regra, foram considerados duplicados os dados cujo registro pudesse ser encontrado mais de uma vez nas tabelas de dimensão, seja for ter sido carregado duas vezes ou por haver duas ou mais linhas iguais na base original.

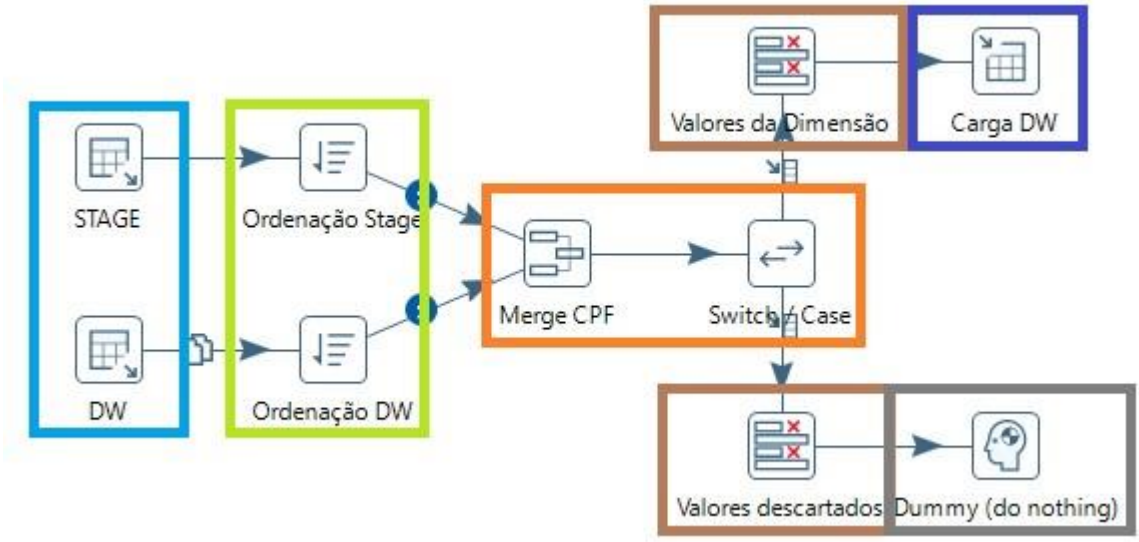

Figura 2: Regra ETL responsável pela remoção de Duplicados.
Fonte: Próprio autor.

A seleção em azul claro indica a entrada de dados de uma tabela da *Stage Area* e de sua tabela correspondente no *Data Warehouse*. Em seguida, na seleção em verde claro, ocorre a ordenação dos dados através de um campo de referência, sendo uma etapa indispensável para a operação seguinte. Após ordenados, os dados passam pelo processo de merge, indicado na seleção laranja. Este merge irá comparar os dados da tabela da *Stage Area* com os da tabela do DW para saber quais são iguais e quais são diferentes. Assim, um identificador será criado, podendo ser `identical` e `new`, respectivamente. Essas duas saídas são enviadas para o Switch/Case que irá direcioná-las de acordo com o que se deseja fazer com cada uma. A saída `identical`, que indica os dados que já estão registrados no DW, irá para o passo indicado em marrom, chamado Valores Descartados, onde os campos selecionados serão descartados na caixa *Dummy*, indicado em cinza. Já a saída `new`, que representa os novos dados, é direcionada para o campo em marrom, Valores da Dimensão, que seleciona os campos que serão enviados para a caixa indicada em azul escuro, responsável por fazer o carregamento dos novos dados na tabela do *Data Warehouse*.

[1] https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-platform/pentaho-dataintegration.html

A padronização dos dados de tipo `string` é de suma importância para a geração de relatórios consistentes e organizados visualmente. Dessa maneira, duas das regras elaboradas para contribuir com este processo são apresentadas na Figura 3.

Figura 3: Regra ETL responsável pela padronização de valores do tipo string.
Fonte: Próprio autor.

As `strings`, dados compostos por letras e outros caracteres, foram reescritas em letra maiúscula e também houve a necessidade de remover acentos. Para tal, foram utilizados dois recursos da ferramenta *Pentaho Data Integration*, onde configura-se os valores de tipo `string` que serão transformados e, então, as operações são realizadas. Esta transformação propicia padrão no aspecto visual dos dados, conforme ilustra a Tabela 1.

TABELA 1. Exemplo de resultado da transformação em string.
Fonte: Próprio autor.

Antes da Transformação	Após a transformação
João	JOAO
Joao	JOAO
júlia	JULIA

RESULTADOS E DISCUSSÃO

Respeitando a estrutura do sistema ERP fonte dos dados operacionais e conhecendo os tipos de dados e seus respectivos formatos de preenchimento, foram elaboradas regras ETL através da ferramenta *Pentaho Data Integration*, que resultaram em dados padronizados e consistentes para que pudessem ser carregados no Data Warehouse. As regras ETL apresentadas neste artigo constituem a normalização do DW, sendo chamadas de regras de limpeza.

Remover dados repetidos faz com que o *Data Warehouse* seja conciso e mantenha as informações que de fato são relevantes. As análises e consultas têm seu tempo de espera reduzido, além de trazerem corretamente os valores solicitados, sem haver possibilidade de redundância ou ambiguidade. Segundo Oliveira, Rodrigues e Henriques (2004), esta regra de transformação é uma das responsáveis pela integridade e consistência dos dados. Assim, ao comparar o volume de dados integrados sem remoção de duplicados com o volume de dados integrados e padronizados pelo DW, a diferença foi significativa, indicando que as consultas haviam sido agilizadas.

A padronização das `strings` aparenta ter caráter opcional no que se refere a análise e geração de relatórios ou gráficos. Entretanto, quanto mais organizados e padronizados visualmente, melhor é interpretação e conclusão a respeito das informações geradas. Além disso, quanto mais normalizados são os dados que compõem um *Data Warehouse*, mais eficazes são seus resultados de análise.

Os resultados obtidos por meio das regras apresentadas garantiram dados uniformes, objetivos e promovem a normalização e qualidade dos dados que virão a ser integrados no DW, conforme a recomendação de Kimball e Caserta (2004).

CONCLUSÕES

O processo ETL foi a base da construção do *Data Warehouse* de Vendas, pois estabeleceu a relação entre a estrutura das diferentes fontes de dados e a padronização dos dados. Como resultado obteve-se uma base de dados com qualidade e pronta para *Business Intelligence* e Aprendizado de Máquina no auxílio ao processo de tomada de decisão. Além disso, o DW de Vendas possibilitou a integração do sistema ERP de Pet Shops com dados padronizados e normalizados.

A partir de um questionário qualitativo, concluiu-se que este trabalho escalou soluções para outros DW em trabalhos futuros. Um exemplo seria um DW para compras de produtos de fornecedores dos lojistas com novas regras de negócio. Pode-se avaliar também o impacto positivo das regras na estruturação dos dados quanto à eficácia de consultas e na análise dos dados; com resultados consistentes sobre as vendas do sistema ERP de Pet Shop.

A análise dos resultados tornou clara o impacto de regras de limpeza sobre os dados, pois a remoção de duplicidades reduziu o volume de dados extraídos, mostrando que o DW está conciso e pragmático, e a padronização de `strings` tornou os resultados de consultas mais organizados e de fácil compreensão. Tendo em vista a importância da consistência dos dados para o apoio às tomadas de decisão, é possível concluir as consultas realizadas com os dados já transformados são capazes de otimizar processos de ERP.

AGRADECIMENTOS

Agradecemos a empresa JN Moura Informática pelo auxílio financeiro e intelectual a este trabalho de pesquisa.

REFERÊNCIAS

ELMASRI, R.; NAVATHE, S. B. Sistemas de Banco de Dados. 6. ed. São Paulo: Pearson Addison Wesley, 2011.

FERREIRA, João et al. O processo etl em sistemas data warehouse. In: INForum. 2010. p. 757-765.

HANLIN, Qin; XIANZHEN, Jin; XIANRONG, Zhang. Research on extract, transform and load (ETL) in land and resources star schema data warehouse. In: 2012 Fifth International Symposium on Computational Intelligence and Design. IEEE, 2012. p. 120-123.

INMON, William H. What is a data warehouse? Prism Tech Topic, v. 1, n. 1, p. 1-5, 1995.

KIMBALL, R.; CASERTA, J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. [S.l.]: Wiley, 2004. 528 p. ISBN 978-0-764-56757-5.

KIMBALL, R.; ROSS, M. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. 3. ed. Indianapolis, USA: John Wiley & Sons Inc., 2013.

OLIVEIRA, Paulo; RODRIGUES, Fátima; HENRIQUES, P. Limpeza de dados-uma visão geral. Data Gadgets, p. 3951, 2004.

RANJAN, V.A comparative study between ETL (Extract, Transform, Load) and ELT (Extract, Load and Transform) approach for loading data into data warehouse. [S.l.], 2009.

SEAH, Boon Keong; SELAN, Nor Ezam. Design and implementation of data warehouse with data model using surveybased services data. In: Fourth edition of the International Conference on the Innovative Computing Technology (INTECH 2014). IEEE, 2014. p. 58-64.

WIJAYA, Rahmadi; PUDJOATMODJO, Bambang. An overview and implementation of extraction-transformationloading (ETL) process in data warehouse (Case study: Department of agriculture). In: 2015 3rd International Conference on Information and Communication Technology (ICoICT). IEEE, 2015. p. 70-74.