VI Encontro de Iniciação Científica e Tecnológica
VI EnICT
ISSN: 2526-6772
IFSP – Câmpus Araraquara
21 e 22 de outubro de 2021

ERP de Pet Shop: técnica ETL na elaboração de um *Data Warehouse* de Compras e Vendas de produtos e serviços

GABRIEL DE ALCANTARA RODRIGUES SOARES [1], RENATA MARIA PORTO VANNI [2,] NATIELLY CRISTINA FRANCISCO [3], TALES BOALIM [4]

[1]Discente do Tecnólogo em Análise e Desenvolvimento de Sistemas, Subsequente Bolsista Fundação, IFSP Câmpus Araraquara, alcantara.g@aluno.ifsp.edu.br
[2]Docente do Instituto Federal de São Paulo – Câmpus Araraquara, rportovanni@ifsp.edu.br
3Analista de Tecnologia da Informação, JN Moura Informática, natielly.francisco@jnmoura.com.br
[4]Gerente de Tecnologias e Inovação, JN Moura Informática, tales@rfidmoura.com.br

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

RESUMO: Este trabalho estendeu um *Data Warehouse* de Vendas de produtos e serviços a clientes de varejo de empresas de Pet Shop através da inclusão de transações de compra de produtos feita pelos varejistas em seus fornecedores. Essa integração aplicou o processo de Extração, Transformação e Carregamento dos dados (*Extract, Transform and Load* - ETL) em um subconjunto de bases operacionais de empresas de varejo no ramo de Pet Shop com dados de transações de compras e vendas de produtos e serviços. Regras de ETL foram criadas e aplicadas nesse subconjunto de bases por meio da ferramenta *Pentaho Data Integration*, efetuando a limpeza, padronização e estruturação dos dados no novo *Data Warehouse* de compras e vendas de produtos e serviços a clientes de varejo de empresas de Pet Shop. Com isso, são apresentadas algumas regras de transformação e padronização para que o *Data Warehouse* extendido tenha os dados normalizados.

PALAVRAS-CHAVE: *Data Warehouse;* ETL; integração de dados; normalização de dados; processo de ETL.

INTRODUÇÃO

Segundo Puga *et al*. (2013), o dado é uma unidade básica de informação que, no cenário de banco de dados, representa um valor e o conjunto desses valores associados a um contexto é tido como sendo uma informação possível de interpretação e análise. A informação é fundamental para produzir conhecimento e tomar decisões, principalmente do ponto de vista do ramo dos negócios. Com isso, um *Data Warehouse* (DW) é tido como uma maneira efetiva para a integração entre os dados, ou seja, para que estejam centralizados e possibilitem análises gerenciais para tomadas de decisões.

Durante a construção de um DW, o processo de Extração, Transformação e Carregamento dos dados (*Extract, Transform and Load* - ETL), que é considerado como sendo a etapa mais crítica e demorada na elaboração de um DW (FERREIRA *et al.*, 2010), é de suma importância porque os dados sofrem um processo de limpeza, normalização e padronização com o foco na qualidade das informações extraídas, tendo em vista também a continuidade do processo, ou seja, a integração no DW a dados futuros. Para isso, os dados são extraídos das fontes originais (por exemplo, bases de dados transacionais), transformados por meio de regras, com foco nas análises futuras e, então, carregados no DW.

Em relação às dificuldades durante o processo de extração de dados, com cada fonte tendo uma característica específica, dependendo do contexto, surge com isso uma complicação para que se tenha clareza durante a compreensão dos dados, pois estão despadronizados. Dessa forma, surge a necessidade da busca por

padronização, limpeza e integração dos dados, com alta demanda de atenção às determinadas particularidades do sistema de origem, respeitando o padrão criado para as tabelas do DW e seus relacionamentos. Além disso, para que o DW tenha sentido para as empresas que o solicitam, é preciso que haja reuniões com a empresa, estudo das fontes de dados para predizer as futuras análises e verificar quais respostas as empresas pretendem obter com estas análises (ERBA, 2020).

Diante do exposto, como uma extensão do projeto "Construção de *Data Warehouse* para Integração, análise e mineração de dados", que, como apresentado na **Fig. 1**, foi voltado para a construção de um DW na empresa JN Moura Informática, houve a elaboração de um novo processo de ETL (ERBA, 2020), análises gerenciais em ferramentas de *Business Intelligence* (BI) (FRANCISCO, 2020) e, por fim, classificações e agrupamentos com mineração de dados (KANEGAE, 2020). O processo foi voltado para vendas no segmento de Pet Shop com dados vindos de um sistema de ERP, enquanto o novo processo é focado nas compras das empresas no mesmo segmento.

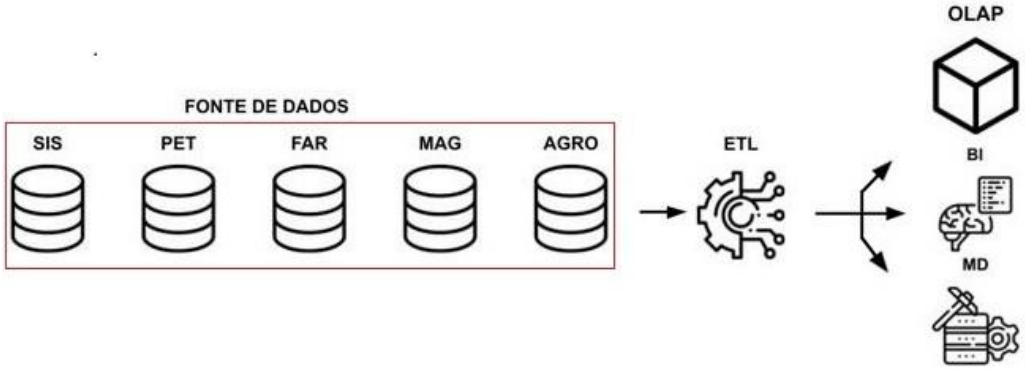

Figura 1. Etapas abrangendo todo o projeto.
Fonte: próprio autor.

FUNDAMENTAÇÃO TEÓRICA

Estudos foram realizados para a contextualização dos conteúdos bases na realização do projeto, no caso, *Data Warehouse*, de modo que fosse possível compreender as maneiras de modelagem, sendo elas: *Star Schema* (Kimball, Ross, 2011) ou *Snowflake Schema* (INMON, 1995). Para ambas as abordagens, o conceito é bem similar, de modo que haja uma (ou várias) tabela fato, que representa o evento a ocorrer (por exemplo, uma compra ou venda) atrelado a pelo menos duas dimensões, que são as características do evento (por exemplo, dados de um fornecedor relacionado a um determinado evento ou dados de um cliente atrelado a esse mesmo evento). As diferenças entre os modelos acontecem na estruturação, pois, no modelo do *Snowflake Schema* (INMON, 1995) pode haver relacionamentos entre as dimensões e outras tabelas que não sejam consideradas tabelas fato, caso que não ocorre no modelo *Star Schema* (Kimball, Ross, 2011).

Além disso, para esta iniciação científica, especificamente, foi estudado o processo de ETL em que, como apresentado na **Fig. 2**, primeiramente, os dados são extraídos de fontes distintas, então, são transformados para que se tenha um padrão adequado através de regras aplicadas (HANLIN et al., 2012) de acordo com o modelo de *Data Warehouse* criado. Após isso, os dados são carregados, primeiramente, à uma zona intermediária chamada *Data Staging Area* (DSA), ou somente *Stage* (FERREIRA *et al*., 2010), que não possui a estrutura do DW, sendo somente tabelas que se assemelham ao modelo de um *Data Warehouse*, porém, sem relacionamento entre elas. Por fim, os dados são carregados da DSA para o DW de forma que eles sejam estruturados de acordo com o modelo adotado, ou seja, tendo os relacionamentos propostos nos modelos teóricos utilizados como base, seja o *Star Schema* (Kimball, Ross, 2011) ou *Snowflake Schema* (INMON, 1995).

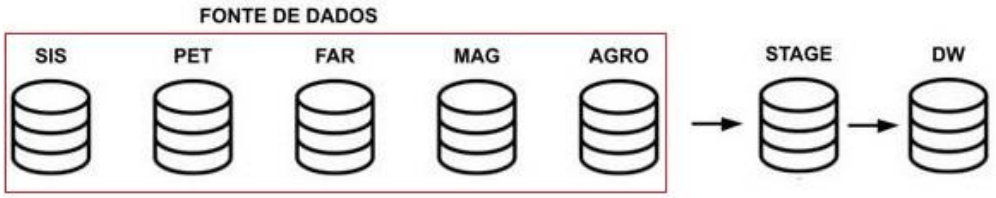

Figura 2. Processo de ETL para construção de um *Data Warehouse*.
Fonte: próprio autor.

METODOLOGIA

Para o desenvolvimento do projeto, a metodologia utilizada se deu pela leitura de livros e artigos sobre *Data Warehouse* e ETL, documentação dos modelos relacionais das bases de dados de Pet Shop, expansão de um modelo de DW contendo as informações mais importantes relacionadas às compras e vendas de uma empresa que agregariam em futuras análises, escolha da ferramenta responsável pelo processo de ETL e, por fim, a elaboração das regras durante o procedimento.

Dentre os processos metodológicos seguidos durante a realização do trabalho de pesquisa, destaca-se a elaboração dos processos de ETL, que foram definidos com o intuito de relacionar 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* desde a extração das fontes operacionais até o carregamento no *Data Warehouse*.

Um dos pontos com maior demanda de atenção durante o processo foram as tuplas duplicadas, ou seja, dados repetidos que poderiam vir a fazer parte do *Data Warehouse*. Com isso, conforme ilustra a **Fig. 3**, foram desenvolvidos passos para que isso não acontecesse.

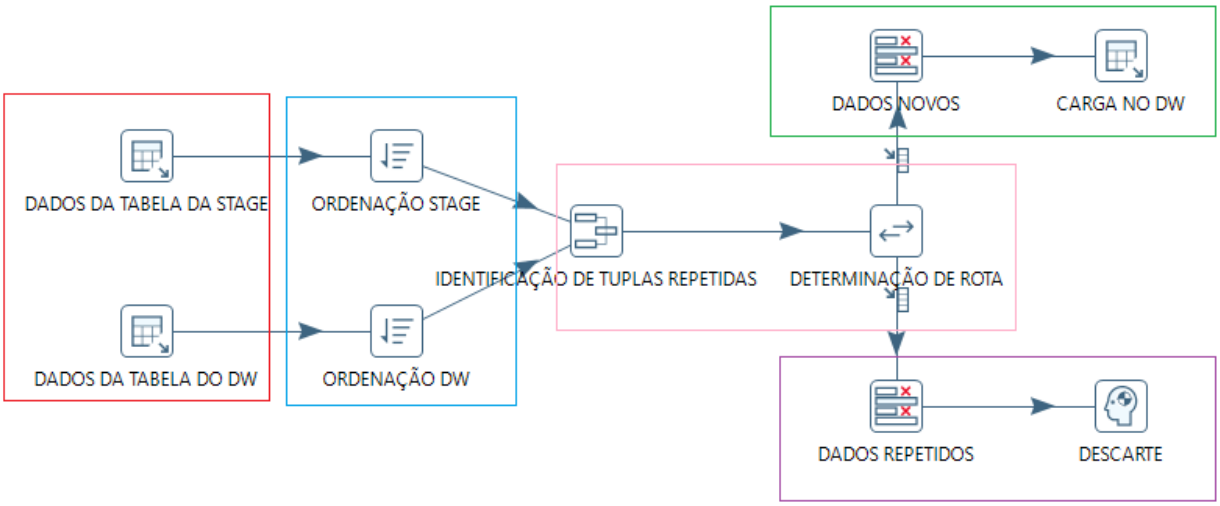

Figura 3. Processo do ETL para que não se tenha dados duplicados no DW.
Fonte: próprio autor.

O trecho destacado em vermelho se refere aos dados vindos de uma determinada tabela da DSA e a sua correspondente no DW, por exemplo, considerando a tabela da *Stage* como sendo Produto, a tabela do *Data Warehouse* seria a Dimensão_Produto. Já o trecho destacado em azul indica uma etapa de ordenação das tuplas através de um determinado atributo especificado, sendo de suma importância para o próximo passo.

Então, após isso, com o trecho destacado em rosa, acontece o processo de identificação das tuplas que são consideradas repetidas com base na condição especificada e, com base nisso, elas podem ter dois destinos, que são especificados no próximo passo, sendo que, as tuplas são consideradas repetidas através de uma pseudocoluna chamada *flagfield* podendo indicar *new*, para valores novos ou *identical* para valores repetidos. Com isso, a etapa destacada em verde indica o processo no caso de novos valores armazenados no DW, ou seja, é feito o carregamento. Por fim, a etapa em roxo indica o processo que ocorre caso os dados sejam duplicados sendo que, na prática, eles são descartados, como indica o *step* nomeado de DESCARTE.

Além do processo de remoção de duplicatas, outro passo fundamental, é a normalização e padronização de *strings*, para isso, ainda utilizando o *Pentaho Data Integration*, é possível remover a acentuação e determinar todos os caracteres para maiúsculo com 2 etapas simples, como é apresentado na Figura 4.

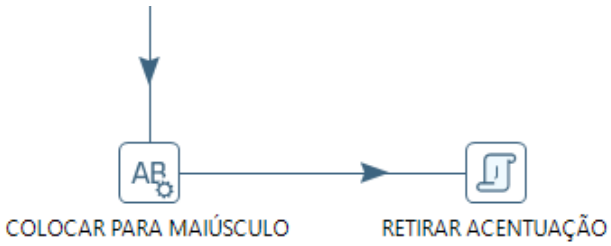

Figura 4. Processo do ETL para normalização de strings que serão carregadas no DW.
Fonte: próprio autor.

RESULTADOS E DISCUSSÃO

Os processos de ETL utilizados durante o projeto tiveram como um dos objetivos transformar as *strings* de modo a padronizá-las e não permitir que houvesse a inserção de tuplas duplicadas durante o carregamento dos dados vindos de bases distintas.

O processo de padronização das strings acontecia em duas etapas:

- Definir todos os caracteres que compõem a *string* para maiúsculo;
- Retirar a acentuação dos caracteres.

Um exemplo do resultado da transformação das *strings* acontece no atributo referente aos nomes das cidades em que três fornecedores, que foram selecionados aleatoriamente, estão localizados, como é apresentado na Tabela 1.

Tabela 1. Comparação entre strings antes e depois do carregamento no *Data Warehouse*

Nomes das cidades em que os fornecedores pertencem vindos das bases transacionais	Nomes das cidades em que os fornecedores pertencem vindos do *Data Warehouse*
São PAuLO	SAO PAULO
Fortaleza	FORTALEZA
Anápolis	ANAPOLIS

Fonte: próprio autor.

A aplicação dos processos de ETL aconteceram primeiramente em dados de teste, ou seja, os que não vinham diretamente do ambiente de produção. Com isso, como fontes dos dados, foram utilizadas compras e vendas, armazenadas em bases de dados, de 53 CNPJs distintos, sendo, no total, referente à 4.348.410 tuplas levadas em consideração durante o processo. Em sua primeira versão, no quesito das quantidades de dados que passavam por cada tabela, os resultados obtidos estão apresentados na Tabela 2.

Tabela 2. Versão 1 da comparação entre as diferenças nas quantidades de dados de acordo com a tabela em relação à etapa do processo de transformação em dados de teste

TABELAS/BASES	BASE TRANSACIONAL	STAGE	*DATA WAREHOUSE*
Empresa	53	53	53
Fornecedor	6336	3956	3956
Nota Fiscal	55895	48953	48953
Tempo	24977	3832	3832
Produto	250655	250655	250655
Compra	579307	573249	573249
Venda	3431187	3391852	3391852

Fonte: próprio autor.

Como é apresentado na Tabela 2, grandes diferenças entre os dados das bases transacionais de Fornecedor, Nota Fiscal, Tempo, Compra e Venda em relação à *Stage* aconteciam, e, para isso, tinham algumas razões:

- Havia, de fato, algumas tuplas duplicadas e, com isso, elas eram descartadas.
- Neste momento, foi utilizado somente a DIM_FORNECEDOR para armazenar os dados de fornecedor. Posteriormente, foi notado que para o mesmo CNPJ de fornecedor presente em duas ou mais empresas distintas, existiam nomes fantasia diferentes cadastrados, com isso, havia um problema na rastreabilidade. Assim, surge a necessidade de ter a tabela EMPRESA-FORNECEDOR, e, com ela, as diferenças entre os dados de fornecedores são reduzidas, como analisado na Tabela 3.
- Os dados de tempo apresentam grande diferença, pois somente um registro de data para cada evento é necessário, com isso, vários são descartados.
- As diferenças em relação às notas fiscais aconteciam por algumas possuírem o valor relativo ao atributo Chave_Nfe, que representa um valor único de nota fiscal eletrônica, como sendo vazio, com isso, havia um problema durante a diferenciação.

Tabela 3. Versão 2 da comparação entre as diferenças nas quantidades de dados de acordo com a tabela em relação à etapa do processo de transformação em dados de teste

TABELAS/BASES	BASE TRANSACIONAL	STAGE	*DATA WAREHOUSE*
Empresa	53	53	53
Fornecedor	6336	5116	5116
Nota Fiscal	55895	55513	55513
Tempo	24977	3832	3832
Produto	250655	250655	250655
Compra	579307	573249	573249
Venda	3431187	3391852	3391852

Fonte: próprio autor.

Na tabela 3, mesmo tendo diferenças, elas não são um problema, pois representam dados duplicados e, para que esse resultado tenha sido atingido, algumas modificações foram feitas:

- Em relação à nota fiscal, os valores que anteriormente vieram vazios, ou seja, sem um valor único para que fossem diferenciados, foram preenchidos com uma concatenação entre o CNPJ da empresa em questão, envolvida na compra, e o código da nota fiscal, que representa um valor inteiro que é auto incrementado, com isso, é garantido que não haja qualquer problema durante a diferenciação das notas fiscais.

- Sobre o fornecedor, após a implementação da tabela EMPRESA-FORNECEDOR, decorrente do problema citado anteriormente, o processo de diferenciar dados de fornecedor, para identificar o que seria considerado repetido e não seria aproveitado, se deu pela comparação entre o CNPJ da empresa em questão juntamente com o código do fornecedor, que é um valor inteiro sequencial. Com isso, todos os dados de fornecedores distintos foram carregados na *Stage* e, posteriormente, no DW.

Após analisar os resultados obtidos com os dados de teste, o processo de ETL foi aplicado com os dados de produção, já com as atualizações comentadas anteriormente, sendo, inicialmente, 45 CNPJs distintos, sendo, no total, referente à 10.470.879 tuplas levadas em consideração durante o processo. Os resultados alcançados estão na Tabela 4.

Tabela 4. Versão 1 da comparação entre as diferenças nas quantidades de dados de acordo com a tabela em relação à etapa do processo de transformação em dados de produção

TABELAS/BASES	BASE TRANSACIONAL	STAGE	*DATA WAREHOUSE*
Empresa	53	47	45
Fornecedor	20606	17243	17243
Nota Fiscal	80455	78155	78155
Tempo	4650	4638	4638
Produto	787458	633519	633519
Compra	856371	815238	815238
Venda	8721286	8395531	8395531

Fonte: próprio autor.

Como apresentado na Tabela 4, durante os processos, os padrões encontrados nas diferenças entre os dados foram os mesmos em relação aos descritos de acordo com a Tabela 2. A única excessão se dá com as empresas, sendo que, da base transacional para a *Stage* há uma redução de 53 para 47, ou seja, até então, dentro do esperado por se tratar de CNPJs repetidos. No entanto, a diferença entre a *Stage* e o DW, que não era esperada, acontece por uma razão:

- Apesar dos 47 CNPJs serem únicos, o que faz com que eles sejam reduzidos para 45 é que algumas empresas estão inativas, com isso, são desconsideradas.

CONCLUSÕES

Com a aplicação dos processos de ETL, os dados foram integrados de maneira concisa, ou seja, sem que haja duplicatas ou *strings* não padronizadas e, além disso, de forma que não haja qualquer tipo de perda de dados durante as etapas, possibilitando, com isso, garantia de qualidade em processos de *Business Intelligence* ou Mineração de dados que utilizem o DW como base.

Além disso, na prática, a remoção de duplicatas durante o processo de ETL faz com que o volume de dados seja menor do que caso não haja esse tipo de filtragem, com isso, o desempenho do DW é melhor. Por fim, com a padronização de *strings* os resultados se tornam mais consistentes fazendo com que análises gerenciais feitas tenham maior relêvância e clareza.

REFERÊNCIAS

ERBA, Ana Guelfi et al. ERP de Pet Shop: técnica ETL na elaboração de um *Data Warehouse* de Vendas. In: **V Encontro de Iniciação Científica e Tecnológica-EnICT (ISSN: 2526-6772)**. 2020.

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

FRANCISCO, Natielly Cristina et al. Análise de Dados utilizando Ferramentas de *Business Intelligence* aplicadas a *Data Warehouse* sobre Vendas de Varejo de Pet Shop. In: **V Encontro de Iniciação Científica e Tecnológica-EnICT (ISSN: 2526-6772)**. 2020.

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.

KANEGAE, Yuri Henri Takashi Queiroz et al. Aplicação de Agrupamento de Dados para a Identificação de Perfis Gerais Entre Clientes. In: **V Encontro de Iniciação Científica e Tecnológica-EnICT (ISSN: 2526-6772)**. 2020.

KIMBALL, Ralph; ROSS, Margy. **The *Data Warehouse* toolkit: the complete guide to dimensional modeling**. John Wiley & Sons, 2011.

PUGA, Sandra et al. Banco de Dados: Implemetação em SQL, PL/SQL e Oracle 11g. **São Paulo, São Paulo**, 2013.