Desenvolvimento de um Data Mart e Automatização do Processo ETL no contexto da Produção Acadêmica do CEULP/ULBRA

Murillo Roseno Feitoza Lima¹, Jackson Gomes¹

¹Ciências da computação - Centro Universitário Luterano de Palmas CEULP/ULBRA
Avenida Teotônio Segurado, 1501 SUL, CEP 77.019-900, Palmas - TO - Brasil

murillo.roseno.lima@gmail.com, jgomes@ceulp.edu.br

Resumo: *Este artigo apresenta o processo de desenvolvimento de um data mart e automatização do processo ETL no contexto da produção acadêmica do CEULP/ULBRA. O procedimento realizado tem como intuito centralizar os dados referentes a produção acadêmica que estão armazenados em diversas fontes de dados da instituição. Deste modo, o presente trabalho tem como finalidade automatizar a integração desses dados a partir do processo ETL e armazená-los em um data mart.*

1. Introdução

O Centro Universitário Luterano de Palmas dispõe de uma base de dados voltada para armazenar documentos e informações acadêmicas, dentre eles artigos, monografias e TCCs, utilizada no software *Biblioteca Digital*, desenvolvido internamente. Os dados armazenados são relacionados a várias outras fontes de dados instituição, que constituem informações relativas às contidas na base de dados da *Biblioteca Digital*, como monografias (documentos), autores e membros de bancas. Essas informações, caso sejam analisadas, tratadas e precisas, podem contribuir para a estratégia de negócio que a instituição possa vir a utilizar.

É nesse cenário que o *data warehouse* está inserido. “Um *data warehouse* (armazém de dados) é uma coleção de dados orientado a assuntos, integrados, variante no tempo e não volátil para suporte ao gerenciamento dos processos de tomada de decisão” (INMON, 1997). Conforme afirmam Elsmari e Navathe (2011), um *data warehouse* é muito distinto de um banco de dados tradicional em sua estrutura, funcionamento e finalidade. O banco de dados tradicional é destinado ao processamento de transações, armazenamento, recuperação e atualização de dados. Por sua vez, o *data warehouse* é projetado exatamente para análise complexa de dados e descoberta de conhecimento. É válido ressaltar que, em comparação com os bancos de dados tradicional, o *data warehouse*, em geral, possui grande quantidade de dados de fontes diversas, integrados e centralizados em um único ambiente.

Um ambiente de *data warehouse* inclui um repositório de dados, mecanismo de processamento e integração de dados. Para realizar o processamento e integração de dados existe um processo denominado *ETL* que, segundo Kimball e Caserta (2004), é a principal etapa na construção de um *data warehouse*. Este processo é composto por três etapas: extração, responsável por extrair os dados; transformação, responsável por realizar a limpeza e as transformações necessárias nos dados; carga, responsável por integrar os dados no ambiente de *data warehouse*.

Nesse contexto o objetivo geral deste trabalho é desenvolver um *data mart* e automatizar o processo *ETL* para centralizar os dados referentes à produção acadêmica, disponíveis na base de dados da biblioteca digital do CEULP/ULBRA. Este artigo apresenta os resultados obtidos e considerações sobre as principais contribuições.

2. Fundamentos

“Um *data warehouse* (armazém) é uma coleção de dados, orientado a um assunto, integrado, tempo-variante e não volátil, para suporte ao gerenciamento dos processos de tomada de decisão” (INMON, 1997, Seção 2.1.1).

“Sua crescente utilização pelas empresas está relacionada à necessidade do domínio de informações estratégicas para garantir respostas e ações rápidas assegurando a competitividade de um mercado altamente competitivo e mutável” (MACHADO, 2004).

A arquitetura do *data warehouse* inclui, além de estrutura de dados, mecanismos para a carga e consulta de dados para o usuário final. Esse conceito define os elementos da arquitetura de um *data warehouse*, como ilustra a Figura 1, uma representação da arquitetura de um *data warehouse*.

Figura 1: Arquitetura de um Data Warehouse

Segundo Machado (2004), a arquitetura define o modelo lógico, independente da estrutura física do *data warehouse*. A arquitetura é constituída por um conjunto de ferramentas que respondem desde a carga até o processamento de consultas, assim como por repositórios de dados, como o *data warehouse* e os *data marts*.

As ferramentas existentes na arquitetura podem ser divididas em dois grupos: a) Ferramentas relacionadas à carga inicial e às atualizações periódicas dos *data warehouse*, ferramentas *ETL* (*Extract, Transform, Load*); e b) Ferramentas relacionadas às consultas orientadas para o usuário final que são responsáveis pela elaboração de relatórios, pesquisas informativas, análise de desempenho e mineração de dados.

As características principais de um Data Warehouse são (IMON, 1997): orientado a assuntos, integrado, variável com o tempo e não volátil.

O *data mart* é um subconjunto de dados de um *data warehouse*, organizado por assunto pertencente a um departamento específico. Inmon (1996) descreve os *data marts* como estruturas de dados que contêm informações de acordo com interesse e necessidade do departamento de uma organização, isto é, as informações são armazenadas por áreas ou assuntos específicos.

“Modelos multidimensionais tiram proveito dos relacionamentos inerentes nos dados para preencher os dados em matrizes multidimensionais, chamadas cubos de dados” (ELMASRI e NAVATHE, 2011). Um cubo é caracterizado por uma tabela de fatos ministrando as dimensões, formado por três elementos básicos: Fatos; Dimensões; Medidas (variáveis).

As tabelas fatos permitem que as medidas sejam visualizadas conforme classificadas nas dimensões. Consequentemente só fazem sentido quando classificadas pelas dimensões as quais os fatos farão sentido e poderão ser analisados. De acordo com Kimball (2002), em uma tabela de fatos, uma linha corresponde a uma medição. Uma medição é uma linha em uma

tabela de fatos. Todas as medições em uma tabela de fatos devem estar alinhadas na mesma granularidade.

Em relação à estrutura ou composição do *Data Mart* pode-se utilizar o Modelo Estrela, que possui um fato ao centro e, ao redor, dimensões ligadas ao fato, formando uma estrela, e o Modelo Floco de Neve que, segundo Machado (2004), é o resultado da decomposição de uma ou mais dimensões com hierarquias entre seus membros.

O processo *ETL* é composto de três etapas: **Extração, Transformação e Carga de dados** (como apresentado na Figura 2). As etapas de extração e carga são obrigatórias para o processo, sendo a transformação/limpeza opcional.

Figura 2: Estrutura do Processo ETL

A etapa de extração abrange a extração de dados dos sistemas de origem e é o primeiro passo do processo *ETL*. Na etapa de transformação são definidas as regras e funções que definem a qualidade dos dados. A etapa de Carga consiste em carregar os dados obtidos da segunda etapa sempre levando em conta a integridade dos dados em relação ao *data warehouse*.

3. Metodologia

Para o desenvolvimento deste trabalho foi utilizado o SGBD (Sistema de Gerenciamento de Banco de Dados) SQL Server 2014 como forma de armazenamento dos dados estruturados do *data mart*.

Para o desenvolvimento e automatização do processo *ETL* foi utilizada a ferramenta *Pentaho Data Integration – PDI*, também conhecido como *Kettle*, é a ferramenta de *ETL* da *Pentaho*. É um software desenvolvido em *Java* baseado no conceito de processo para extração, transformação e carga (*ETL*) de dados (PENTAHO, 2019). O *Pentaho Data Integration (PDI)* fornece ferramentas que incluem *ETL* e agendamento em um ambiente unificado - a interface do cliente *PDI* (PENTAHO, 2019).

A Figura 3 apresenta a estrutura de dados, diagrama Entidade-Relacionamento (ER), da biblioteca digital.

Figura 3: Estrutura de dados, diagrama ER, da biblioteca digital

Os procedimentos adotados são ilustrados pela Figura 4.

Figura 4: Procedimentos adotados no desenvolvimento do trabalho

As etapas ilustradas pela Figura 4 são:

1. Análise do modelo de dados Entidade-Relacionamento do banco de dados da biblioteca digital do CEULP/ULBRA, a fim de extrair o entendimento necessário para a criação de um modelo abstrato do *data mart*;
2. Definição do modelo abstrato do *data mart*, criado a partir dos elementos básicos da modelagem multidimensional (dimensões, fatos, métricas e modelo estrela);
3. Implementação do modelo abstrato, criando assim o modelo físico para armazenar os dados do *data mart*;
4. Integração dos dados, utilizando Kettle; e
5. Criação de *jobs*, responsáveis pela automatização do processo *ETL*.

4. Resultados

O intuito deste trabalho foi o desenvolvimento de um *data mart* e automatização do processo *ETL* para centralizar os dados referentes à produção acadêmica disponíveis nos bancos de dados do CEULP/ULBRA, apresentado suas estruturas e elementos.

O modelo utilizado neste trabalho foi o *Star Schema*, um esquema de dados *Star* consiste em uma tabela central conhecida como a tabela de fatos, rodeada por tabelas de dimensões de forma que a tabela fato esteja ligada as dimensões, formando uma *Star*.

A Figura 5 apresenta o modelo lógico do *data mart* de acordo com o modelo *star schema*, modelo arquitetado com base nas informações levantadas a partir do modelo de dados Entidade-Relacionamento: composta por uma tabela Fato – Publicação e três dimensões: Curso, Pessoa e Tempo.

Figura 5: Modelo lógico do data mart – Modelagem Star Schema

O nome da pessoa, tipo, sexo e as dimensões têm informações descritivas sobre os atributos do modelo, a exemplo de pessoa, curso e tempo. De forma que a tabela fato tenha indicadores(métricas) do modelo como quantidade de publicações.

O processo *ETL* utilizando Kettle/Spoon é apresentado pela Figura 6.

Figura 6: Processo ETL no Kettle/Spoon

Do ambiente apresentado na Figura 6, pode-se destacar contornado pela cor laranja e com a marcação de número 1 o *canvas* do Kettle/Spoon usado para projetar e construir transformações e tarefas para o processo *ETL*, e inspecionar e gerenciar os dados da transformação. A transformação da dimensão tempo apresentada constitui uma rede de etapas com um fluxo de trabalho de dados que executa suas tarefas *ETL* como: leitura, validação, refino, transformação e gravação dos dados com destino a tabela de dimensão Tempo do *data mart*. O mesmo processo ETL apresentado se repete para as demais dimensões de Curso e Pessoa.

Para a carga da tabela fato Publicação foi inserido ao processo a tarefa de consulta/atualização da dimensão, permitindo a pesquisa de valores em uma dimensão. A Figura 7 ilustra o processo de carga da tabela fato Publicação.

Figura 7: Processo de carga da tabela fato Publicacao no Kettle/Spoon

Apesar da Figura 7 apresentar a carga da tabela fato Publicacao, a diferença entre a carga de uma tabela dimensão é que os dados que foram armazenados foram dados indicativos e não descritivos.

A automatização do processo *ETL* foi feita com *job* no Kettle/Spoon como mostra a Figura 8.

Figura 8: Automação do processo ETL com job no Kettle/Spoon

O *job* é semelhante ao fluxo de trabalho do processo *ETL*, a Figura 8 apresenta o modelo do *job* criado para automatizar o processo *ETL* coordenando recursos de execução de atividades de *ETL*. O *job* iniciou com a execução da transformação *ETL* da dimensão tempo, assim que foi concluída houve um salto para a execução da próxima dimensão, o processo se repetiu para a dimensão de pessoa e para o fato Publicação, com término na etapa *success*.

5. Conclusões

O presente trabalho teve como propósito o desenvolvimento de um *data mart* e automação do processo *ETL* para centralizar os dados referentes à produção acadêmica disponíveis nos bancos de dados do CEULP/ULBRA. Para a realização do trabalho foi necessário o aprendizado de conceitos de desenvolvimento de *Data Warehouse*, *Data Mart* e *ETL*, assim como ferramentas necessárias para a conclusão do objetivo do trabalho.

Para a automação do processo *ETL* a partir do Kettle pode-se destacar algumas conclusões sobre a utilização da ferramenta, possui uma interface gráfica intuitiva para projetar e gerenciar o processo *ETL*, sem a necessidade de utilizar recursos de programação por código. Atende uma variedade de conexões com diferentes fontes de dados de origem e destino. A documentação referente a cada tarefa do processo de transformação do *ETL* facilita o entendimento do uso da ferramenta. A semelhança entre o modelo de fluxo de trabalho do processo *ETL* e da criação de *jobs* ajudam na execução e coordenação.

Referências

- INMON, W. H. The data warehouse and data mining. Communications of the ACM, v. 39, n.11, p.49-50, 1996.
- _____, W. H.; HACKATHORN, R. D. Como usar o Data Warehouse. IBPI, 1997.
- _____, W. H. Como construir o data warehouse. Indianapolis: Wiley Publishing Inc., 1997.
- ELMASRI, Ramez. Sistemas de banco de dados. 6 ed, São Paulo: Pearson Addison Wesley, 2011
- KIMBALL, Ralph; ROSS, Margy. The Data Warehouse Toolkit. Rio de Janeiro: Editora Campus, ed. 2, p. 494, 2002.
- _____, Ralph; CASERTA, Joe. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Indianapolis: Wiley Publishing, Inc., p. 467, 2004.
- MACHADO, Felipe Nery Rodrigues. Tecnologia e Projeto de Data Warehouse: Uma visão multidimensional. São Paulo: Editora Érica Ltda, 2004.
- PENTAHO, Open Source Business Intelligence. Disponível em: <http://kettle.pentaho.com/>
Acesso em: 29 mar. 2019.