CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS
Recredenciado pela Portaria Ministerial nº 1.162, de 13/10/16, D.O.U nº 198, de 14/10/2016
ASSOCIAÇÃO EDUCACIONAL LUTERANA DO BRASIL
Murillo Roseno Feitoza Lima

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

Palmas – TO
2019

Murillo Roseno Feitoza Lima

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

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Trabalho de Conclusão de Curso II (TCC II) do curso de bacharel em Ciência da Computação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. M.e Jackson Gomes Souza.

Palmas – TO

2019

Murillo Roseno Feitoza Lima

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

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Trabalho de Conclusão de Curso II (TCC II) do curso de bacharel em Ciência da Computação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. M.e Jackson Gomes Souza.

Aprovado em: _____/_____/_____

BANCA EXAMINADORA

_____

Prof. M.e Jackson Gomes Souza

Orientador

Centro Universitário Luterano de Palmas – CEULP

_____

Prof. M.e Fabiano Fagundes

Centro Universitário Luterano de Palmas – CEULP

_____

Prof. M.e Parcilene Fernandes de Brito

Centro Universitário Luterano de Palmas – CEULP

Palmas – TO

2019

RESUMO

LIMA, Murillo Roseno Feitoza. **Desenvolvimento de um data mart e automatização do processo etl para centralizar os dados referentes à produção acadêmica dispóniveis nos bancos de dados do ceulp/ulbra.** 2019. Trabalho de Conclusão de Curso (Graduação) - Curso de Ciência da Computação, Centro Universitário Luterano de Palmas, Palmas/TO, 2019.

Este trabalho apresenta o processo de desenvolvimento de um *data mart* e automatização do processo ETL no contexto da produção acadêmica do CEULP/ULBRA, a fim de, centralizar os dados referentes disponíveis na Biblioteca Digital. Neste trabalho é apresentado as características inerentes de um *data warehouse*, tipos de arquiteturas e implementação de *data marts*, o conceito de granularidade de dados para a contextualização além da explicação da modelagem de um Data Warehouse esclarecendo os componentes do processo, tabela fato, tabela dimensão, tipos de medidas, modelo estrela e modelo floco de neve e o processo ETL, a partir dos passos de extração, transformação e carga dos dados. A metodologia é composta por etapas abordando os materiais utilizados, sendo para o armazenamento de dos o uso do SGBD Microsoft SQL Server 2014 e para o desenvolvimento e automatização do processo ETL o *Pentaho Data Integration* e os procedimentos necessários para o desenvolvimento do projeto, partindo da análise do modelo ER da biblioteca digital, definição do modelo abstrato do *data mart*, criação do modelo físico do *data mart*, integração de dados e criação de *jobs*. Por meio de análises feitas no modelo Entidade-Relacionamento, foi possível identificar e elaborar o modelo lógico e físico do *data mart*, base para as demais etapas do procedimento utilizado. Por fim, os resultados obtidos foram a estrutura física do *data mart* implementada para receber os dados, a tabela fato e as tabelas dimensões criadas, o processo ETL e a criação de *Jobs* para a automatização desses processos foram criadas.

Palavras-chave: *data warehouse*; *data mart*; ETL; integração de dados; biblioteca digital; *pentaho*; *PDI*; *kettle*; *jobs*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

"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, 1990).

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*.

Conforme afirmam Elmasri 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 tradicionais, o *data warehouse*, em geral, possui grande quantidade de dados de fontes diversas, integrados e centralizados em um único ambiente, Elmasri e Navathe (2011).

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. Os dados armazenados são relacionados a várias outras fontes da mesma instituição, que constituem informações relativas às contidas na base de dados da Biblioteca Digital como documentos, autores e membros de bancas.

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. Neste sentido objetiva-se especificamente:

- definir os artefatos, o modelo dimensional e fazer a modelagem do *data mart*;
- implementar o *data mart*;
- automatizar o processo *ETL*.

O trabalho está estruturado da seguinte forma: o capitulo 2 apresenta o referencial teórico dividido entre *Data Warehouse*, que apresenta definições e técnicas para implementação, e Processo ETL que apresenta definições. No capítulo 3 são apresentados a metodologia, os materiais e procedimentos utilizados no desenvolvimento deste trabalho. O capitulo 4 apresenta os resultados obtidos e como foi o processo de desenvolvimento do *data*

mart e do processo ETL para as dimensões e fatos, e como foi a automatização desse processo com a utilização de *job*. No capítulo 5 são relacionadas as conclusões deste trabalho, e por fim as referências.

2 REFERENCIAL TEÓRICO

Esta seção tem como objetivo apresentar e descrever conceitos e definições necessários para o desenvolvimento do trabalho acerca dos assuntos de *Data Warehouse, Data Mart*, Processo *ETL*.

Assim, esta seção está estruturada da seguinte forma: na subseção 2.1 são apresentados os conceitos de *Data Warehouse*, como as suas principais características, elementos e fundamentos; subseção 2.2 apresenta os conceitos do processo *ETL*, sua abrangência e funcionamento.

2.1 *Data Warehouse*

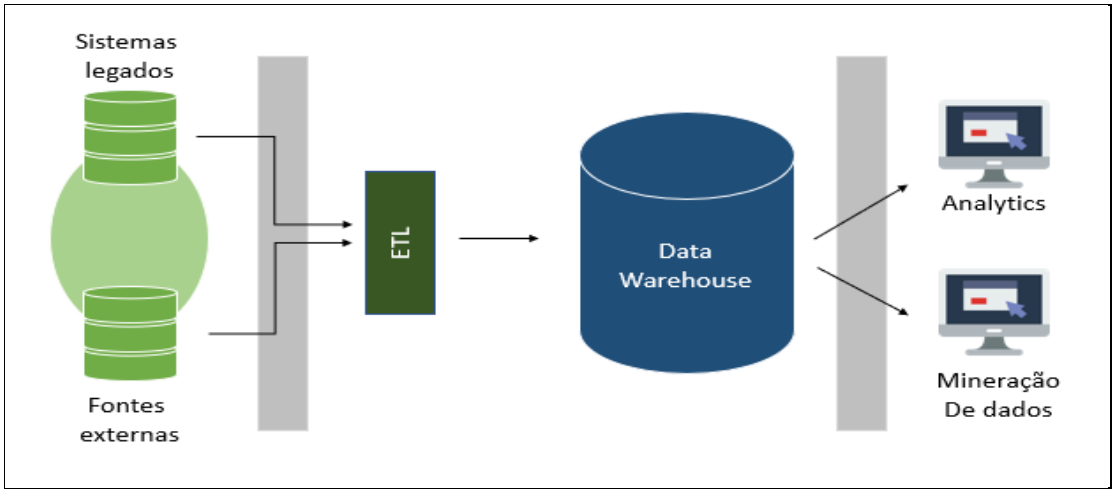
"Um *data warehouse* (armazém) é uma coleção de dados, orientado a um assunto, integrado, tempo-variante e não volátil, para suporte ao gerenciamento dos processos de tomada de decisão" (INMON, 1997, Seção 2.1.1).

Inmon (1997) define um *data warehouse* através de suas principais características, já Machado, relata a visão através da perspectiva de uma empresa, abordando sua visão de negócio, quando afirma que a "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).

De acordo com Elmasri e Navathe (2011) é necessário que haja um modelo de dados apropriados para um *data warehouse*, ao contrário dos bancos de dados transacionais que oferecem acesso de dados disjuntos e normalmente heterogêneos, um *data warehouse* é um depósito de dados integrados de múltiplas fontes, processados para armazenamento em um modelo multidimensional.

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.*



Fonte: Adaptado de Machado (2004)

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 mart*s.

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, Seção 2.2).
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.

Quanto aos repositórios, o *data warehouse* funciona como um grande conjunto de todos os dados, enquanto os *data mart*s são subconjuntos do *data warehouse*, direcionados para uma visão específica de um conjunto de dados.

A escolha da arquitetura leva em consideração fatores relativos a infra-estrutura disponível e recursos para investimento. A seção 2.1.1 apresenta as características inerentes de um *data warehouse*.

2.1.1 Características inerentes de um *data warehouse*

Inmon (1997) define quatro características para um *data warehouse*: orientados a assuntos, integrados, variáveis com o tempo e não volátil. Estas características serão mais detalhadas nas subseções a seguir:

2.1.1.1 Orientados a assuntos

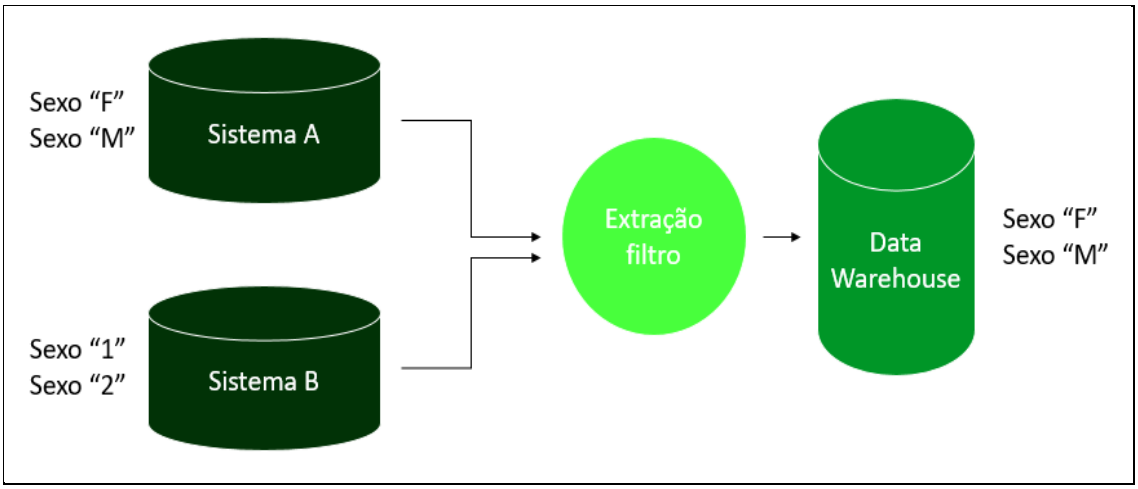
A orientação por assunto é uma característica marcante de um *data warehouse*. Os dados e processos de interesse são agrupados por assunto. Em contraste aos bancos de dados tradicionais que são voltados a processos e transações operacionais.

De acordo com Gilmar (2002) assunto é um conjunto de informações relativas à determinada área estratégica de uma empresa. Por exemplo: em um comércio de vendas, as áreas estratégicas poderiam ser os produtos, os revendedores, as contas, os clientes, dentre outras.

2.1.1.2 Integrado

Compreende a conexão de integração entre os dados. É por seu intermédio que os dados de diversos sistema convergem em torno de um assunto (dado a característica de orientação por assunto), nesse processo também há a padronização e unificação do que será visualizado pelo usuário. A Figura 2, apresenta a idealização do conceito.

Figura 2. Idealização do conceito de Integração.



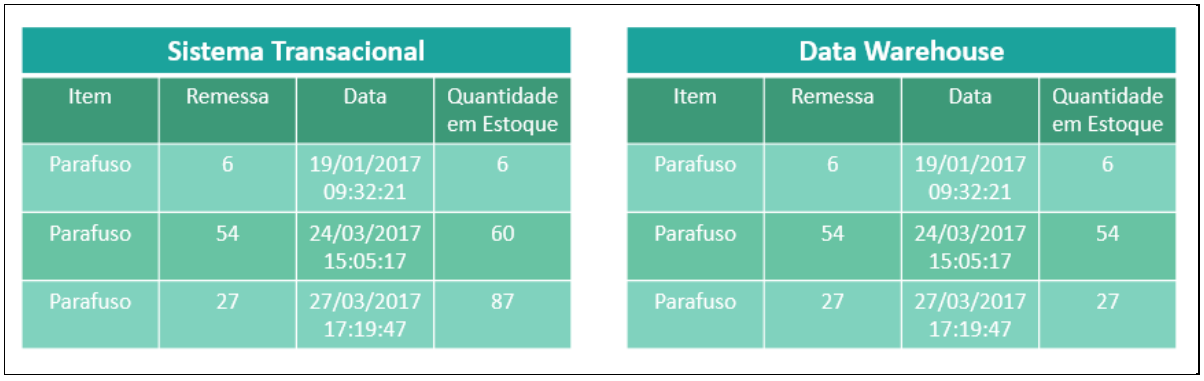
Fonte: Adaptado de Machado (2004)

A Figura 2, apresenta um exemplo clássico, comumente citado por vários autores, que é a informação do sexo. No sistema A é apresentado com F para feminino, M para masculino. Já no sistema B, ocorre exatamente o oposto, com um para feminino e dois para masculino. Para ocorrer a integração entre os sistemas, é essencial a adotar um padrão único para sexo, por exemplo, F, M.

2.1.1.3 Variável com o tempo

Segundo Inmon (1996), os *data warehouses* são variáveis em função ao tempo, o que corresponde a manter o histórico de dados durante um espaço de tempo superior aos existentes em sistemas transacionais. Em um ambiente transacional os dados são constantemente atualizados, portanto qualquer operação de consulta refletirá a verdade para o sistema no momento da operação. No *data warehouse*, o dado refere-se a um tempo específico, significando que não é atualizável, logo a cada ocorrência deste dado uma nova entrada é criada para marcar esta mudança. A Figura 3 apresenta o comportamento de ambos ambientes, no cenário de quantidade de itens em um estoque de uma loja de materiais para construção.

Figura 3. Comportamento de um sistema transacional e de um *data warehouse*.



Sistema Transacional

Item	Remessa	Data	Quantidade em Estoque
Parafuso	6	19/01/2017 09:32:21	6
Parafuso	54	24/03/2017 15:05:17	60
Parafuso	27	27/03/2017 17:19:47	87

Data Warehouse

Item	Remessa	Data	Quantidade em Estoque
Parafuso	6	19/01/2017 09:32:21	6
Parafuso	54	24/03/2017 15:05:17	54
Parafuso	27	27/03/2017 17:19:47	27

Fonte: Adaptado de Machado (2004)

A Figura 3, apresenta o comportamento de um dado específico no ambiente do sistema transacional e o mesmo dado em um ambiente de *data warehouse*, consolidando a visão de funcionamento de ambos ambientes em função do tempo. No sistema transacional a quantidade em estoque do item parafuso é atualizada em função da remessa. Em contradição, no *data warehouse*, a cada remessa é criada uma entrada, para marcar a mudança.

2.1.1.4 Não volátil

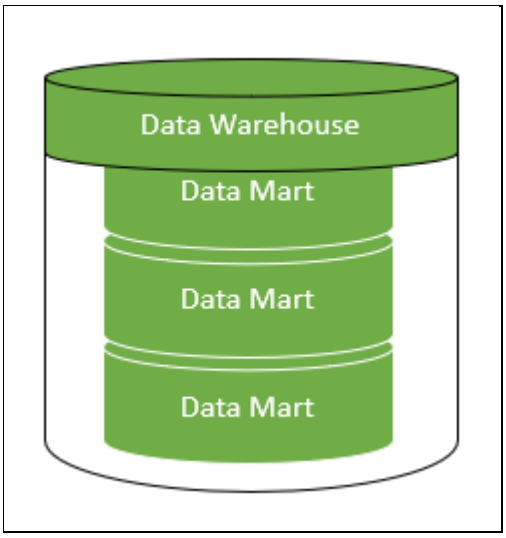
Um *data warehouse* não volátil deve permitir duas operações básicas: uma carga inicial dos dados e a carga dos dados através da incremental. A carga dos dados no *data warehouse* ocorre através de blocos de registros e não registro a registro como ocorre em bancos de dados tradicionais, os quais realiza-se diversas validações a cada registro.

De forma geral, foram apresentadas informações referentes as características intrínsecas de um *data warehouse*, nas seções a seguir serão abordados conceitos sobre *data mart* e componentes que formam a estrutura multidimensional desses armazenamentos de dados.

2.1.2 *Data Mart*

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. A Figura 4, apresenta o conceito de um conjunto de *data mats* dentro de um *data warehouse*.

Figura 4. Conjunto de *Data Mart* dentro de um *Data Warehouse*.



Fonte: Machado (2004, p. 44)

A seção a seguir apresentará os tipos de arquitetura e formas de implementações de um *data mart*.

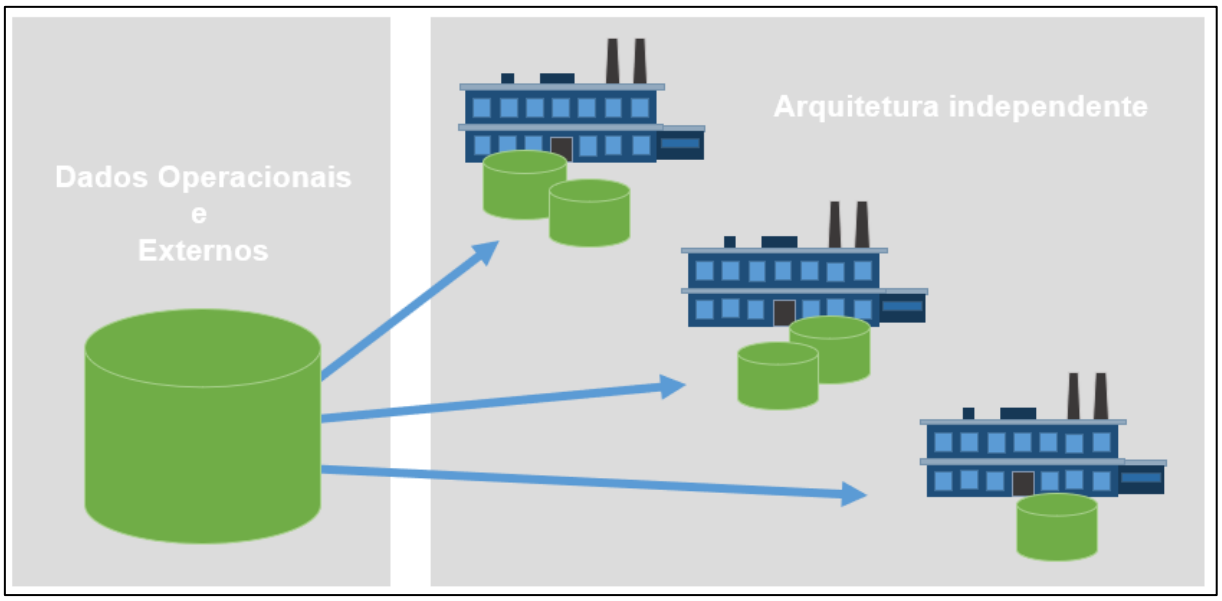
2.1.2.1 Tipos de Arquitetura e de Implementação de *Data Mart*

A seleção de uma arquitetura determinará ou será determinada pelo local onde o *data warehouse* ou *data mart*s estarão residindo. Assim a forma de implementação é influenciada por fatores como a infra-estrutura, recursos tecnológicos, a arquitetura escolhida, o escopo da implementação e principalmente pela necessidade ou não de acesso corporativo dos dados, assim como pelo retorno de investimento desejado e velocidade de implementação (MACHADO, 2004).

2.1.2.1.1 Arquitetura de *Data Mart* Independente

Na arquitetura independente, os dados são controlados para atender as necessidades especificas de um departamento, sem foco corporativo, este fato implica que as informações de um *data mart* não terão nenhuma conexão com outro *data mart* de outros departamentos do negócio. Machado (2004) afirma que este tipo de arquitetura resulta em rápida implementação, em consequência de baixo impacto nos recursos tecnológicos. A Figura 5 apresenta uma arquitetura de *data mart* independente.

Figura 5. Arquitetura de *Data Mart* Independente.

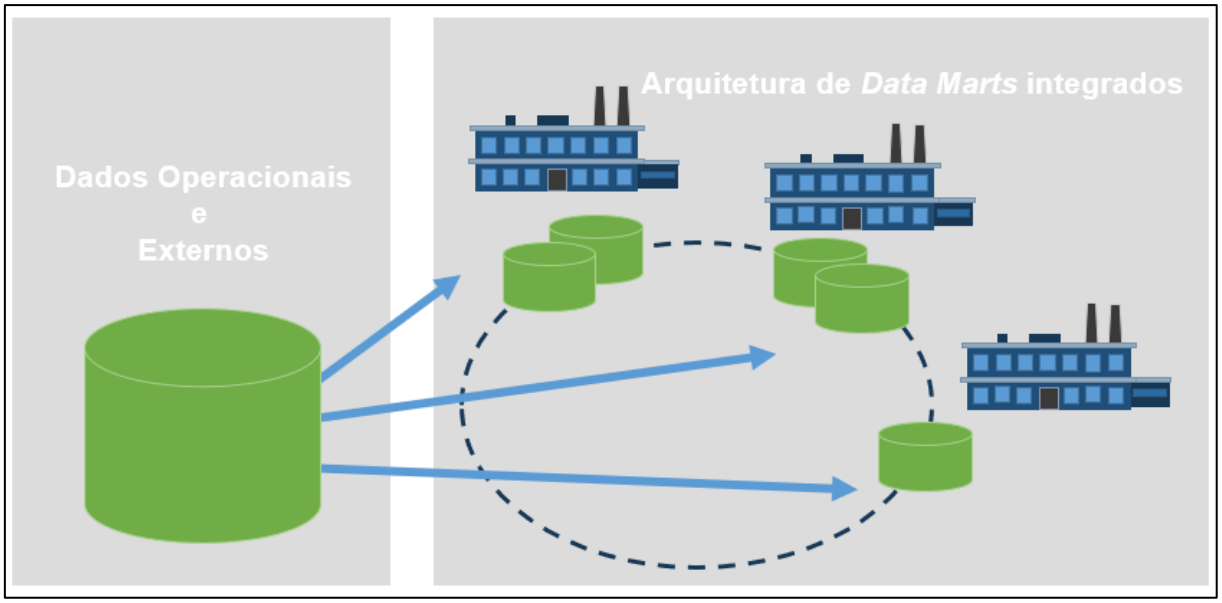


Fonte: Machado (2004, p. 50)

2.1.2.1.2 Arquitetura de *Data Mart*s Integrados

A arquitetura de *data mart*s integrados é o próprio *data warehouse* distribuído em múltiplos *data mart*. Apesar de os *data mart*s serem implementados separadamente por grupos de trabalho ou departamentos, os dados serão integrados ou interconectados, permitindo a visão dos dados provenientes por todos os *data mart*. Dando assim uma visão corporativa maior dos dados e informações, em consequência elevando o tempo de desenvolvimento e o nível de complexidade de requisitos (MACHADO, 2004). A Figura 6 apresenta uma arquitetura de *data mart* integrado.

Figura 6. Arquitetura de *Data Mart* Integrado.



Fonte: Machado (2004, p. 51)

2.1.2.1.3 Implementação *Top Down*

A implementação *top down* requer planejamento, sua abrangência envolve todos os departamentos que farão parte do projeto. Antes do processo se iniciar deve ser considerado alguns fatores sobre o ambiente, conforme Machado (2004) indica, em torno das informações sobre o ambiente transacional que será trabalhado, das fontes de dados que serão utilizadas, segurança, estruturas de dados, qualidade de dados a ser considerada, padrões de dados e modelos de dados. Deste modo, a Tabela 1 apresenta algumas características do modelo *top down*.

Tabela 1 - Vantagens e Desvantagens da Implementação *Top Down*.

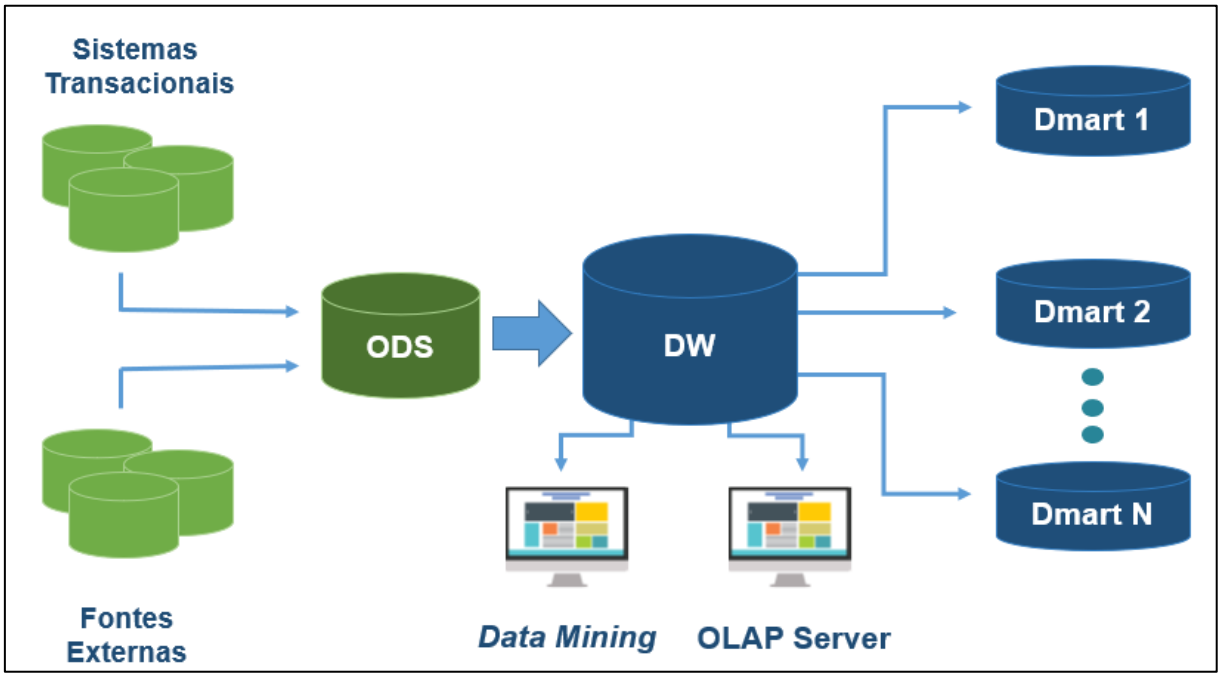
Vantagens	Desvantagens
Herança de arquitetura: arquitetura e dados reutilizados do *DW*, permitem uma fácil manutenção.	**Implementação muito longa:** dificulta a garantia de apoio político e orçamentário.
Visão de empreendimento: o *DW* concentra todos os negócios da empresa, sendo possível extrair dele níveis menores de informações.	**Alta taxa de risco:** não existem garantias para o investimento nesse tipo de ambiente.
Repositório de metadados centralizado e simples: facilidade na manutenção de metadados.	**Heranças de cruzamentos funcionais:** é necessária uma equipe de desenvolvedores e usuários finais altamente capacitados.

Controle e centralização de regras: garante a existência de um único conjunto de aplicações para extração, limpeza e integração dos dados, além de processos centralizados de manutenção e monitoração.	**Expectativas relacionadas ao ambiente:** a demora do projeto e a falta de retorno podem induzir expectativas nos usuários.

Fonte: Adaptado de Machado (2004, p. 53-54)

A Figura 7, apresenta o modelo de implementação *top down* para um *data mart* integrado.

Figura 7. Modelo de Implementação *Top Down* em *Data Mart* Integrado.



Fonte: Machado (2004, p. 53)

2.1.2.1.4 Implementação *Bottom Up*

Este tipo de implementação permite o desenvolvimento por departamentos diferentes sem a necessidade te se ter uma infraestrutura definida, visando o desenvolvimento incremental para o *data warehouse* conforme a implementação dos *data mart* independentes. Este processo conforme Machado (2004), dificulta garantir a padronização da modelagem dimensional, podendo ocorrer redundâncias de dados e inconsistências entre os *data mart*s. Desta forma, a Tabela 2 apresenta as vantagens e desvantagens do modelo botton-up.

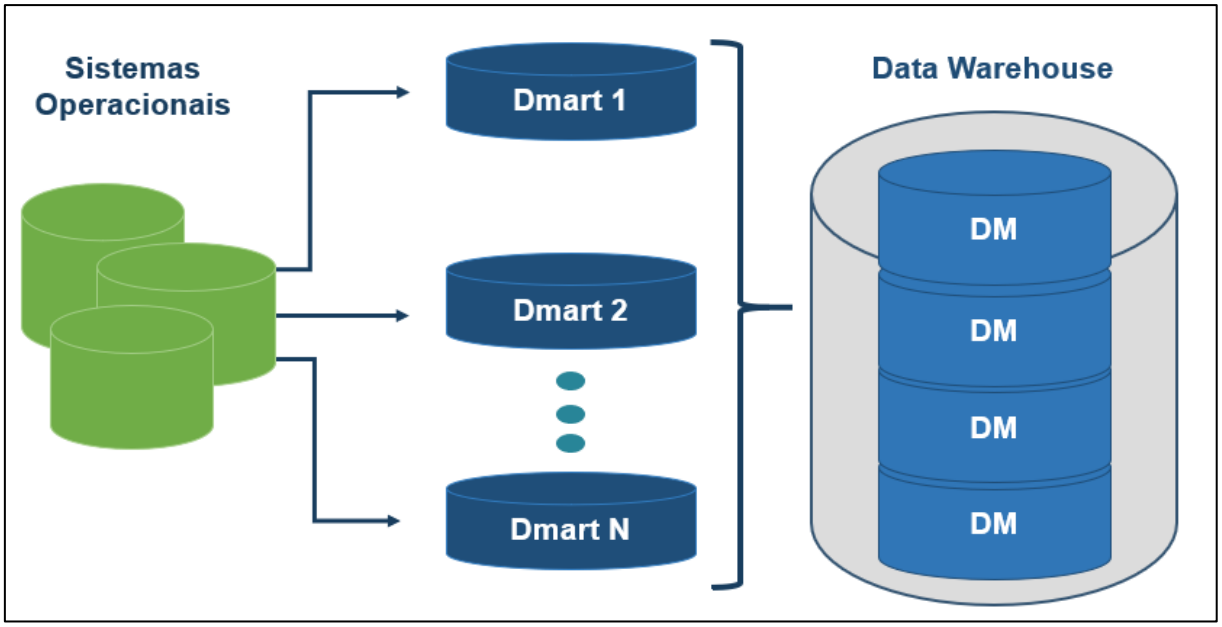
Tabela 2 - Vantagens e Desvantagens da Implementação *Bottom Up*.

Vantagens	Desvantagens
Implementação rápida: desenvolvimento direcionado, reduz o tempo de construção.	**Perigo de *legamarts*:** um dos maiores perigos na criação de *DM* independentes é acabar sendo transformados em sistemas legados, que dificultam e inviabilizam futuras integrações. Tornando-se problema e não solução.
Retorno rápido: arquitetura incremental baseada em *DM*, permite uma base para investimentos adicionais, elevando a confiança.	**Desafio de possuir a visão de empreendimento:** durante a construção dos *DMs* incrementais, é necessário que se mantenha um rígido controle do negócio como um todo.
Manutenção do enfoque da equipe: permite que as principais áreas de desenvolvimento sejam priorizadas pela equipe.	**Administrar e coordenar múltiplas equipes e iniciativas:** em caso de desenvolvimento de *DM* em paralelo, é necessário um rígido controle a fim de evitar perda da padronização.
Herança incremental: reduz o risco do projeto, devido ao crescimento da equipe a medida que as entregas incrementais acontecem.	**A maldição do sucesso:** ao término de um *DM* os usuários felizes querem mais informações para o seu projeto. Enquanto os próximos terão que aguardar o incremento. Sendo necessário que a equipe vença desafios políticos.

Fonte: Adaptado de Machado (2004, p. 55)

A Figura 8, apresenta o modelo de implementação *bottom up* para um *data mart* independente.

Figura 8. Modelo de Implementação *Bottom Up* em *Data Mart* Independente.



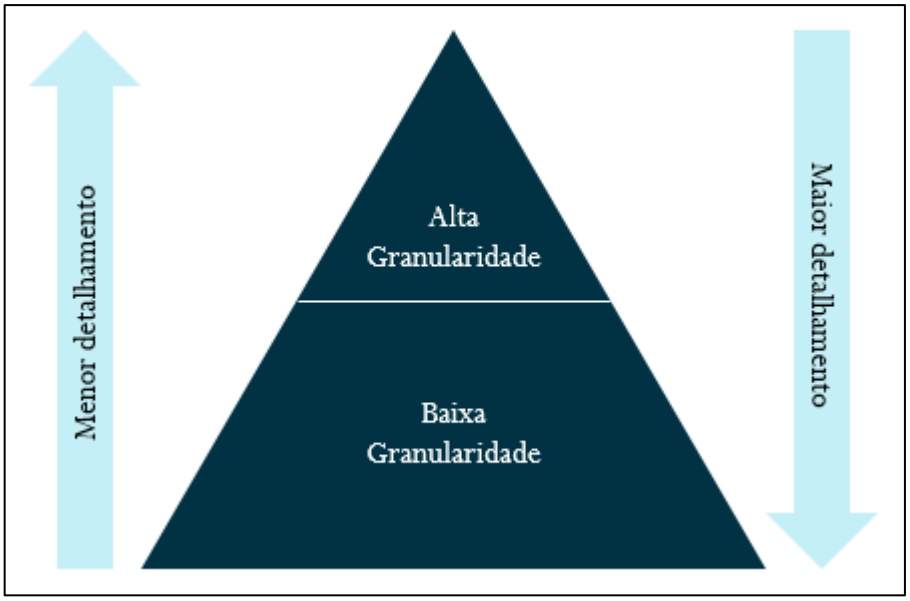
Fonte: Machado (2004, p. 53)

2.1.3 Granularidade

Segundo Machado (2004), a granularidade de dados refere-se ao nível de sumarização dos elementos e de detalhe disponíveis nos dados. Quanto mais detalhado for o *data warehouse*, mais baixo será o nível de granularidade, pois o *DW* terá mais detalhes nos dados e por consequência mais volume. Portanto, quanto menos detalhes existir, maior será o nível de granularidade e, consequentemente, menor será o volume de dados. Este fato é relevante porque afeta diretamente a performance do *data warehouse*, já que envolverá o processamento de um volume maior ou menor de dados dependendo da granularidade definida.

É válido ressaltar que a escolha do nível de granularidade é feita a partir de análise detalhada das necessidades de um projeto. O nível de detalhes de dados indicará a performance e o tamanho do armazenamento do *data warehouse*. A Figura 9, apresenta a relação intrínseca entre nível de detalhe e volume de dados.

Figura 9. Relação entre o nível de granularidade e o volume de dados.



Fonte: Adaptado de Machado (2004)

A Figura 9, acima, apresenta de forma visual a relação entre o nível de granularidade e o volume de dados. O nível de granularidade afeta diretamente o volume de dados armazenados em um *data warehouse* ou *data mart*. A granularidade é importante, pois servirá como parâmetro para o tipo de consulta aplicada ao *data warehouse*.

2.1.4 Modelagem do *Data Warehouse*

"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).

A modelagem multidimensional é uma técnica de concepção e visualização de um modelo de dados de um conjunto de medidas que descrevem aspectos comuns de negócios. É utilizada especialmente para sumarizar e reestruturar dados e apresentá-los em visões que suportem a análise dos valores desses dados.

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. Para tanto as subseções a seguir apresentam características dos três elementos básicos da modelagem multidimensional e as técnicas de modelagem dimensional.

2.1.4.1 Tabela Fato

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.

Toda decisão significativa tem como fundamento um fato, sobre o qual se analisa as pequenas dimensões comportamentais, relacionadas a ele. Sendo essencial deter amplo conhecimento para pôr em prática sua decisão.

Para a identificação e entendimento de um fato é preciso primeiro reconhecer os quatro pontos de referência de um fato, denominados de pontos cardeais de um fato. Um fato consegue, numa única palavra, transmitir quatro informações: Onde aconteceu o fato; quando aconteceu o fato; quem executou o fato; O que é o objetivo do fato.

Machado (2004), trata fato como uma coleção de itens de dados, composta de dados de medidas e de contexto. A característica básica de um fato é que ele é representado por valores numéricos e implementado em tabelas denominadas tabelas de fato. A granularidade de uma tabela de fatos é a definição do que constitui um único registro da tabela de fatos.

“A tabela de fatos é a principal tabela de um modelo dimensional, onde as medições numéricas de interesse da empresa estão armazenadas” (KIMBALL, 2002).

A idealização de uma tabela de fato pode ser exemplificada com uma guia de atendimento idealizada para planejamentos de eventos de formaturas, conforme Tabela 3.

Tabela 3 - Tabela Fato

VENDA
Identificador (PK)
Cliente (FK)
Instituicao (FK)
Evento (FK)
Data
Valor

Fonte: Adaptado de Machado (2004)

A Tabela 3, apresenta informações contidas em uma tabela fato, como identificadores e valores numéricos referentes ao fato Venda.

2.1.4.2 Tabela Dimensão

Dimensões são os elementos que participam de um fato, assunto de negócios. A característica da dimensão é determinar o contexto dos assuntos do negócio. Dimensões não possuem atributos numéricos, pois são somente descritivas e classificatórias dos elementos que participam de um fato. Dimensões são componentes vinculados a algum fato. Exemplos de dimensões: Cliente, Produto, Fornecedor e Tempo.

2.1.4.3 Tipos de Medidas

As medidas ou métricas são os atributos numéricos de um fato. Para Machado (2004), é o elemento que permiti demonstrar o desempenho de um indicador de negócios relativo às dimensões que compõem um fato. Exemplos de medidas são: quantidade vendida, receita bruta, lucro líquido.

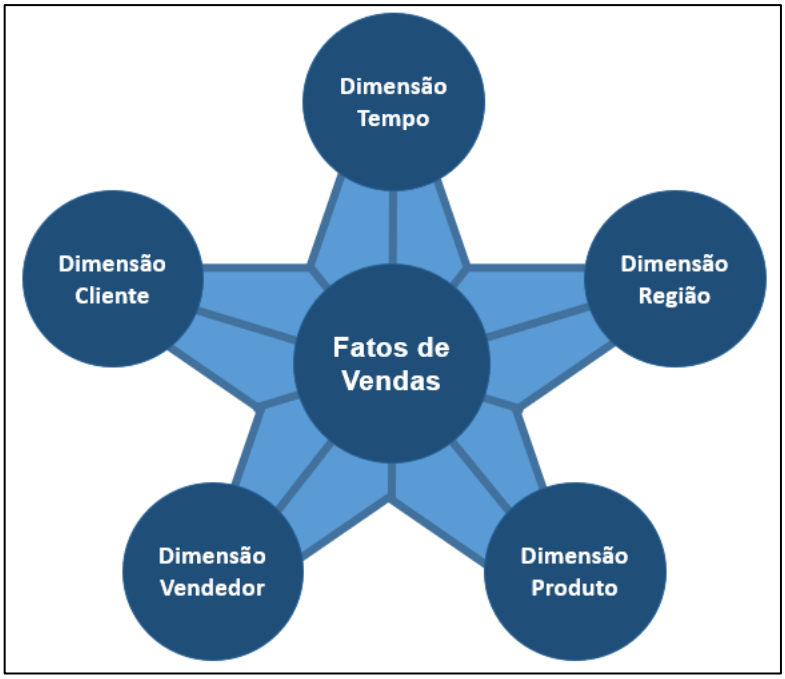
As medidas se classificam em dois tipos:

a) Valores aditivos: são aqueles referentes ao fato sobre os quais podem ser aplicadas operações de matemática básica, somo, subtração e média. Normalmente representado por valores, quantidades e ocorrência.
b) Valores não aditivos: referentes aos fatos que não podem ser manipulados livremente, como valores percentuais ou relativos. Representam os indicadores de desempenho do fato.

2.1.4.4 Modelo Estrela (Star Schema)

O modelo estrela (ou *star schema*) possui um fato no centro, e ao redor estão dispostas diversas dimensões ligadas ao fato, formando uma estrela. A Figura 10 apresenta o modelo estrela.

Figura 10. Modelo Estrela.



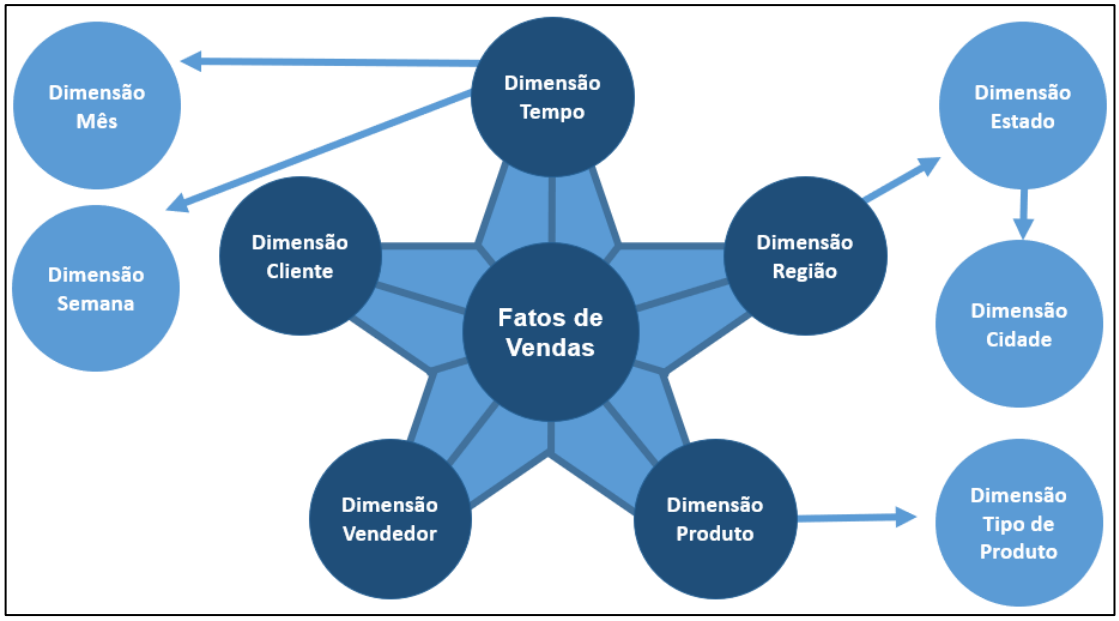
Fonte: Machado (2004, p. 93)

Não existe a normalização nas entidades de dimensões, Machado (2004), os relacionamentos entre a entidade fato e as dimensões acontece através de simples ligações entre as duas tabelas, em um relacionamento de um para muitos no sentido da dimensão para o fato.

2.1.4.5 Modelo Floco de Neve (*Snowflake*)

O modelo multidimensional floco de neve (ou *snowflake*) é semelhante a uma estrela, possui um fato no centro da estrela e as dimensões a sua volta. A Figura 11 apresenta o modelo floco de neve.

Figura 11. Modelo Floco de Neve.



Fonte: Machado (2004, p. 94)

Ocorre a terceira forma normal aplicada nas entidades de dimensões, evitando redundância de dados. Segundo Machado (2004), o modelo floco de neve é o resultado da decomposição de uma ou mais dimensões com hierarquias entre seus membros.

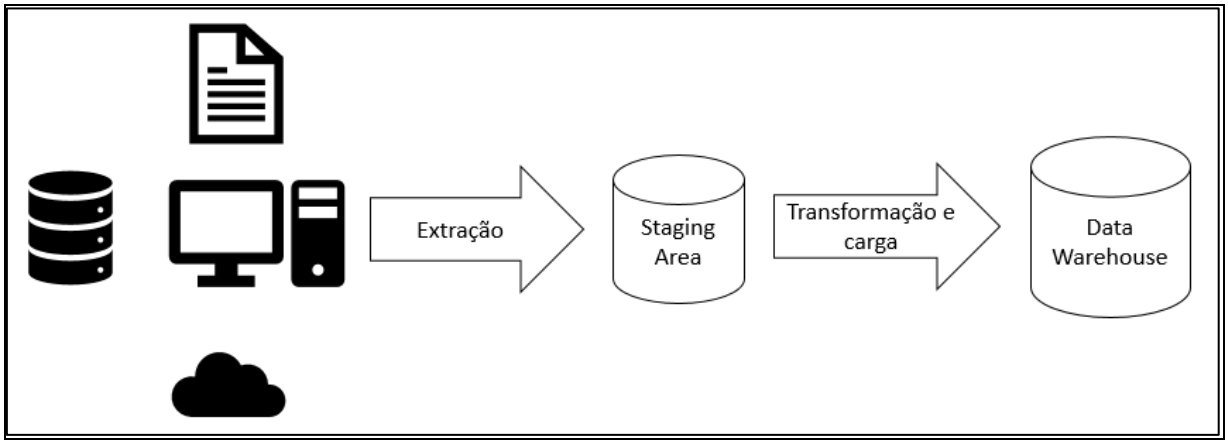
2.2 Processo *ETL*

> "Um sistema *ETL* devidamente projetado extrai dados de sistemas de origens, impõe padrões de qualidade de dados e consistência, conforma dados para que fontes independentes possam ser usadas em conjunto e, finalmente, entrega em um formato que os usuários finais possam usar para tomar decisões." (KIMBALL; CASERTA, 2004).

As particularidades do processo *ETL* apresentadas por Kimball e Carseta (2004, p.xxi): Remove erros e corrige dados ausentes; fornece medidas documentadas de confiança nos dados; Captura o fluxo de dados transacionais para proteção; ajusta dados de múltiplas fontes a serem usadas juntas; e estrutura os dados a serem utilizados pelas ferramentas do usuário final.

O processo *ETL* é composto de três etapas: **Extração**, **Transformação** e **Carga de dados** (como apresentado na Figura 12). As etapas de extração e carga são obrigatórias para o processo, sendo a transformação/limpeza opcional.

Figura 12. Estrutura do *ETL*.



Fonte: Adaptado de Kimball e Caserta (2004)

As seções a seguir apresentam de forma mais detalhada as etapas do processo *ETL* que foram apresentadas na Figura 12.

2.2.1 Extração de dados

A etapa de extração abrange a extração de dados dos sistemas de origem e é o primeiro passo do processo *ETL*. Cada fonte de dados tem seu conjunto distinto de características que precisam ser gerenciadas para efetivamente extrair dados para o processo *ETL*.

Cada dado a ser extraído pode estar em diferentes sistemas (diferentes fontes de dados), "As informações podem ser encontradas nas mais variadas fontes, como, por exemplo, em bancos de dados relacionais, em arquivos-textos, arquivos binários, planilhas eletrônicas, banco de dados orientado a objeto etc." (GONÇALVES, 2003, p. 66).

Ao realizar a etapa de extração certifique-se de ter a documentação apropriada para que o processo seja compatível logicamente e fisicamente com suas políticas, procedimentos e padrões de *ETL* estabelecidos. (KIMBALL; CASERTA, 2004, p. 56).

Na etapa de extração é necessário planejar o processo a ser aplicado, assim deve-se desenvolver o Mapa de Dados Lógico (comumente conhecido como o relatório da linhagem de dados) para mapear as informações relevantes, a fim de, utilizá-lo na implementação física da etapa. Conforme Kimball e Caserta (2004, p. 58), o documento contém a definição de dados para os sistemas de origem do *data warehouse*, o modelo de dados do *data warehouse* de destino e a manipulação exata dos dados necessários para transformá-lo do seu formato original para o seu destino. Na Figura 13, pode-se observar os componentes presentes no Mapa de Dados Lógico.

Figura 13. Mapa de Dados Lógico.

Alvo					Fonte				Transformação
Nome da Tabela	Nome da Coluna	Tipo de Dados	Tipo de Tabela	Tipo de SCD	Nome do Banco de Dados	Nome da Tabela	Nome da Coluna	Tipos de Dados	

Fonte: Adaptado de Kimball e Caserta (2004)

A Figura 13 ilustra os tipos de dados específicos que representam um mapa de dados lógico, Kimball e Caserta (2004) descrevem esses componentes como apresentado a seguir:

a) **Alvo da tabela de destino:** o nome físico da tabela como aparece no *data warehouse*.
b) **Alvo da coluna de destino:** o nome da coluna na tabela do *data warehouse*.
c) **Tipo de dados:** o tipo de dado assumido pelo campo.
d) **Tipo de tabela:** indica se a tabela é um fato, dimensão ou subdimensão.
e) **Tipo de SCD (dimensão lentamente variável):** para dimensões, o componente aborda um tipo indicado podendo ser 1, 2 ou 3. Este indicador pode variar para cada coluna na dimensão. Por exemplo, dentro da dimensão cliente, o último nome pode requerer o comportamento de Tipo 2 (reter o histórico), enquanto o primeiro nome pode requerer Tipo 1 (substituição).
f) **Banco de dados de origem:** o nome da instância do banco de dados onde reside os dados da fonte. Este componente geralmente é a sequência de conexão necessária para se conectar ao banco de dados. Também pode ser o nome de um arquivo como aparece no sistema de arquivos. Nesse caso, o caminho do arquivo também deve ser incluído.
g) **Nome da tabela de origem:** o nome da tabela onde os dados de origem se originam. Haverá muitos casos em que mais de uma tabela é requerida. Nesses casos, basta listar todas as tabelas necessárias para preencher a tabela relativa no *data warehouse* de destino.
h) **Nome da coluna de origem:** a coluna ou colunas necessárias para preencher o alvo. Basta listar todas as colunas necessárias para carregar na coluna alvo. As associações das colunas fontes estão documentadas na seção de transformação.
i) **Transformação:** a manipulação exata requerida dos dados de origem, por isso corresponde ao formato esperado do alvo. Este componente geralmente é anotado em *SQL* ou pseudo-código.

Depois de ter bem definido e planejado o processo de extração, o passo seguinte é a transformação dos dados, que é apresentada na seção seguinte.

2.2.2 Transformação dos dados

Esta etapa é a fase subsequente à extração dos dados. No caso, nesta fase não só é criado as regras de conversão de acordo com os padrões estabelecidos, como também transforma e faz a limpeza dos dados. A correção de erros de digitação, a descoberta de violações de integridade, a substituição de caracteres desconhecidos, as padronizações de abreviações podem ser exemplos desta limpeza (GONÇALVES, 2003).

A transformação é onde são definidas as regras e funções que definem a qualidade dos dados. Ajustes são realizados devidamente nos dados extraídos, os dados são copiados para ***staging area***, onde é realizado o processo de transformação dos dados, através de diversos tratamentos, para determinar a qualidade de dados.

Segundo Kimball e Caserta (2004, p. 115), as características mais relevantes para garantir a qualidade de dados são:

a) **Precisão:** os dados não podem perder suas características originais, os valores e descrições descrevem seus objetivos associados de forma sincera e fiel. Por exemplo, o nome da cidade em que o autor vive atualmente é chamada Palmas. Portanto, dados precisos sobre esse endereço residencial precisam conter Palmas, pois o nome da cidade está correto.
b) **Unicidade:** os valores e as descrições nos dados podem ser tomados para ter apenas um significado. Por exemplo, existem pelo menos duas cidades no Brasil chamadas Palmas, mas há apenas uma cidade no Tocantins chamada Palmas. Portanto, dados precisos sobre um endereço nesta cidade precisam conter Palmas como o nome da cidade e Tocantins como o nome do estado para ser inequívoca. Evitando assim duplicações de informação.
c) **Consistência:** os valores e as descrições nos dados usam uma convenção de notação constante para transmitir o seu significado, ou seja, os fatos devem apresentar consistência com as dimensões que o compõem. Por exemplo, o estado do Tocantins pode ser expresso em dados como TO, ou Tocantins. Para ser consistente, os dados precisos sobre os endereços residenciais atuais devem utilizar apenas uma convenção (como o nome completo do Tocantins) para nome de estado e cumpri-lo.
d) **Completo:** existem dois aspectos de completude. O primeiro é garantir que os valores e descrições individuais dos dados são definidos (não nulos) para cada instância, por exemplo, garantindo que todos os registros que devem ter endereços atuais realmente façam. O segundo aspecto garante que o número agregado de

> registros está completo ou garante que você de alguma forma não perdeu registros em algum lugar em seu fluxo de informações.

Um dos tratamentos necessários é manter a integridade dos dados através da limpeza e filtragem por meio de programas ou rotinas que tentam encontrar irregularidades e resolvê-las, garantindo assim a consistência uma característica prezada pela qualidade de dados.

2.2.3 Carga dos dados

A última etapa do processo *ETL* é a de carga de dados para o novo ambiente. Consiste, em carregar os dados obtidos da segunda etapa sempre levando em conta a integridade dos dados em relação ao *data warehouse*. Após uma carga inicial, são necessárias rotinas de atualizações, para propagar as modificações ocorridas nos sistemas fontes para o *data warehouse*.

Os dados a serem entregues ao *data warehouse* são as tabelas de dimensões e tabelas de fatos. “As tabelas de dimensões fornecem o contexto para as tabelas de fatos e, portanto, para todas as medições apresentadas no *data warehouse*.” (KIMBALL; CASERTA, 2004, p. 161).

As tabelas de dimensões fornecem os pontos de entrada para o *data warehouse*, assim, possuindo mais relevância do que as tabelas de fatos. Uma dimensão é comumente definida como um grão, sendo a fonte de um conjunto de dados. O grão estabelece exatamente o que representa uma única linha da tabela de fatos. Segundo Kimball e Ross (2013, p, 39), “o grão deve ser declarado antes de escolher as dimensões ou fatos porque cada dimensão ou fato do candidato deve ser consistente com o grão”, está consistência impõe quais dados são capturados por um determinado processo de negócio.

Por sua vez, as tabelas de fatos mantêm as medidas de uma empresa. Uma medida é uma quantidade determinada pela observação, se uma medida existe então pode ser tomada como um fato.

3 METODOLOGIA

Esta seção apresenta a metodologia do trabalho, englobando: a metodologia utilizada para chegar ao objetivo do trabalho e a descrição das ferramentas que serão utilizadas para o desenvolvimento do *data mart*.

3.1 Materiais

No desenvolvimento deste trabalho, as tarefas relacionadas ao armazenamento de dados utilizaram o SGBD (Sistema de Gerenciamento de Banco de Dados) *Microsoft SQL Server*, versão 2014.

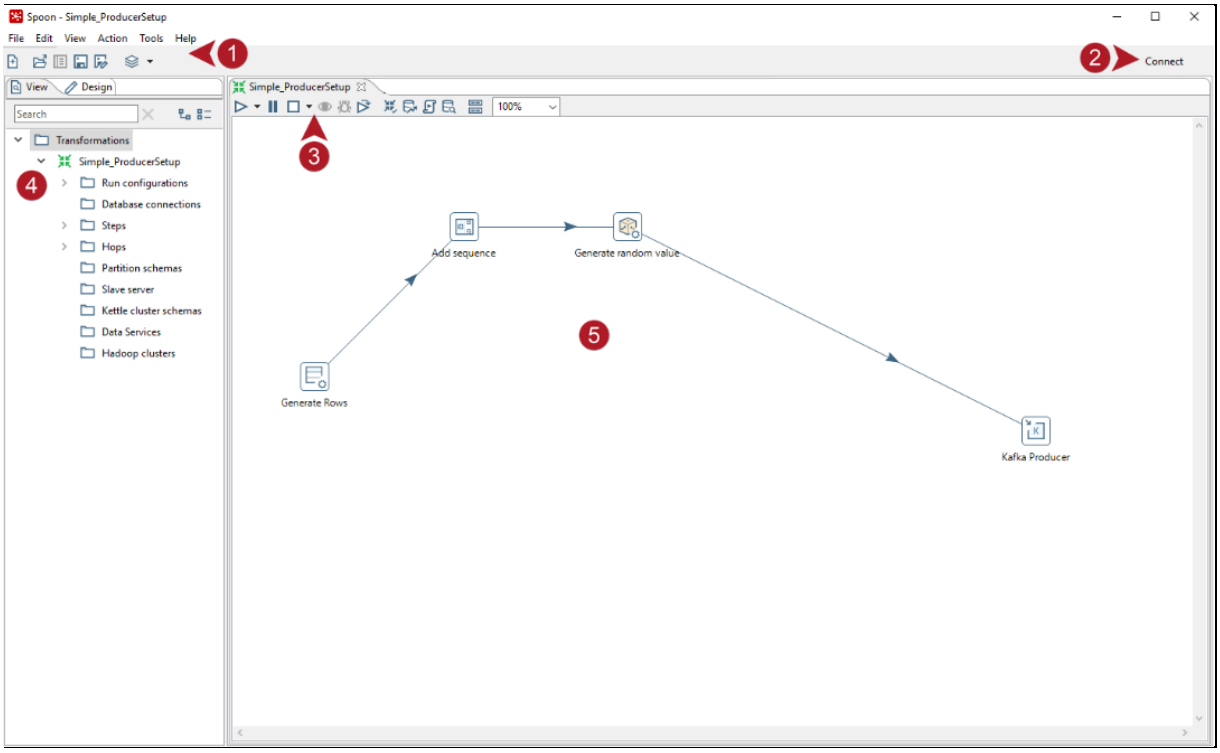
Para o desenvolvimento e automatização do processo *ETL* foi utilizada a ferramenta *Pentaho Data Integration* (PDI). Também conhecida como *Kettle*, é a ferramenta de *ETL* da *Pentaho*. É um software desenvolvido em *Java* baseado no conceito de processo para extração, transformação e carga (ETL) de dados (PENTAHO, 2019).

Assim o *Pentaho* (*Kettle*) é um conjunto de componentes e ferramentas que permitem movimentação de dados, integração de dados não estruturados e manipulações de dados através de múltiplas fontes. Os principais componentes do *Pentaho Data Integration* são (PENTAHO, 2019):

a) ***Spoon*:** uma ferramenta de interface gráfica do usuário utilizada para projetar e gerenciar um processo de transformações ETL (PENTAHO, 2019). Desempenha as funções de fluxo de dados típicos como a leitura, validação, refino, transformação, gravação de dados em uma variedade de diferentes fontes de dados e destinos;
b) ***Kitchen*:** uma ferramenta de linha de comando que permite executar tarefas;
c) ***Pan*:** uma ferramenta de linha de comando que permite executar transformações; e
d) ***Carte*:** um servidor *web* que permite a execução remota de transformações e *Jobs*.

O *Pentaho Data Integration* fornece ferramentas que incluem ETL e agendamento em um ambiente unificado – a interface do cliente PDI (PENTAHO, 2019). Esse ambiente integrado permite que o usuário utilize uma ferramenta gráfica onde o trabalho não necessita da escrita de código. A Figura 14 demonstra um modelo de gerenciamento de processo no *Kettle*.

Figura 14. Modelo *PDI/Kettle.*



A Figura 14 apresenta a interface gráfica do *kettle*(*Spoon*), este ambiente permite a criação e projeção de transformações e tarefas ETL, além de agendar quando as tarefas devem ser executadas.

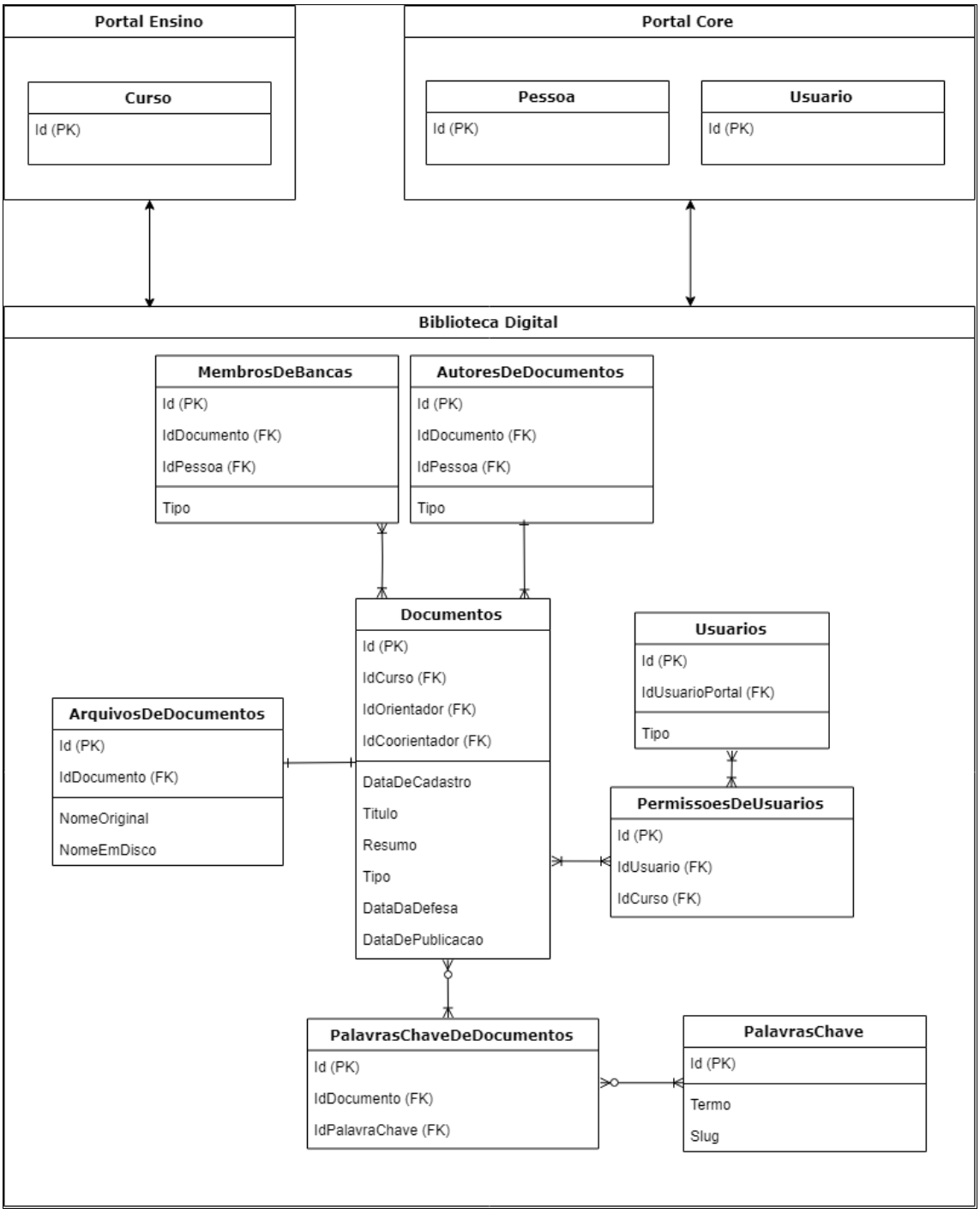
O ambiente é divido em 5 marcações, a marcação de número 1 indica a barra de acesso a ações comuns, a marcação de número 2 mostra o botão de conexão para repositórios de armazenamento de transformações e tarefas ETL, a marcação de número 3 indica a barra de ferramentas para executar ações de transformação ou de *job*, a marcação de número 4 mostra o painel que contém as guia de design, fornece uma lista de etapas ou entradas que são usadas para criar transformações e tarefas, e a guia de exibição, fornece informações, como conexão de bancos de dados e etapas usadas para as transformações e tarefas, a marcação de número 5 indica o *canvas*, usado para projetar e construir transformações e tarefas para as atividades de ETL.

De acordo com a (PENTAHO, 2019), a perspectiva de trabalho no cliente *PDI* é a de projetar tarefas *ETL* e transformações, além de agendamentos de *jobs* e transformações.

Além do *SQL Server* e do *Pentaho* este trabalho utilizou o banco de dados relacional da Biblioteca Digital do CEULP/ULBRA. Esse banco de dados armazena documentos e informações acadêmicas como artigos, monografias e TCCs. Há também outros bancos de

dados em utilização na instituição que servem de fonte para dados relacionados no banco de dados da Biblioteca Digital: **PortalCore** (contém dados de pessoas e usuários do sistema) e **ProtalEnsino** (contém dados de cursos). A Figura 14 apresenta o modelo relacional desses bancos de dados.

Figura 15. Modelo Relacional do banco de dados da Biblioteca Digital e dos bancos de dados PortalCore e PortalEnsino.



O modelo relacional apresentado pela figura indica as tabelas (tabela 4 a 14) e seus campos, apresentados em detalhes a seguir.

Tabela 4 – BibliotecaDigital.Documentos

BibliotecaDigital.Documentos	
Id	Chave primária; o identificador do documento
IdCurso	Chave estrangeira para **PortalEnsino.Curso**, referencia **Id**; representa o curso associado ao documento
IdOrientador	Chave estrangeira para **PortalCore.Pessoa**, referencia **Id**; representa o professor que orienta o trabalho
IdCoorientador	Chave estrangeira para **PortalCore.Pessoa**, referencia **Id**; representa o coorientador do trabalho
DataDeCadastro	Data e hora; data e hora de cadastro do documento
Titulo	Texto; título do documento
Resumo	Texto; resumo do documento
Tipo	Texto; tipo do documento
DataDaDefesa	Data e hora; data e hora da defesa do documento
DataDePublicacao	Data e hora; data e hora de publicação do documento

Tabela 5 – BibliotecaDigital.MembrosDeBancas

BibliotecaDigital.MembrosDeBancas	
Id	Chave primária; o identificador do membro da banca
IdDocumento	Chave estrangeira para **BibliotecaDigital.Documentos**, referencia **Id**; representa o documento associado ao membro da banca
IdPessoa	Chave estrangeira para **PortalCore.Pessoa**, referencia **Id**; representa o professor que é membro da banca do trabalho
Tipo	Texto; tipo do membro da banca

Tabela 6 – BibliotecaDigital.AutoresDeDocumentos

BibliotecaDigital.AutoresDeDocumentos	
Id	Chave primária; o identificador do autor do documento
IdDocumento	Chave estrangeira para **BibliotecaDigital.Documentos**, referencia **Id**; representa o documento associado ao autor
IdPessoa	Chave estrangeira para **PortalCore.Pessoa**, referencia **Id**; representa a pessoa que é autora do trabalho
Tipo	Texto; tipo do autor do documento

Tabela 7 – BibliotecaDigital.ArquivosDeDocumentos

BibliotecaDigital.ArquivosDeDocumentos	
Id	Chave primária; o identificador do arquivo do documento
IdDocumento	Chave estrangeira para **BibliotecaDigital.Documentos**, referencia **Id**; representa o documento associado ao arquivo
NomeOriginal	Texto; nome original do documento
NomeEmDisco	Texto; nome do documento em disco

Tabela 8 – BibliotecaDigital.PalavrasChaveDeDocumentos

BibliotecaDigital.PalavrasChaveDeDocumentos	
Id	Chave primária; o identificador da palavra-chave do documento
IdDocumento	Chave estrangeira para **BibliotecaDigital.Documentos**, referencia **Id**; representa o documento associado a palavra-chave
IdPalavraChave	Chave estrangeira para **BibliotecaDigital.PalavrasChave**, referencia **Id**; representa a palavra-chave associada ao documento

Tabela 9 – BibliotecaDigital.PalavrasChave

BibliotecaDigital.PalavrasChave	
Id	Chave primária; o identificador da palavra-chave
Termo	Texto; termo da palavra-chave
Slug	Texto; slug da palavra-chave

Tabela 10 – BibliotecaDigital.PermissoesDeUsuarios

BibliotecaDigital.PermissoesDeUsuarios	
Id	Chave primária; o identificador da permissão do usuário
IdUsuario	Chave estrangeira para **BibliotecaDigital.Usuarios**, referencia **Id**; representa o usuário associado a permissão
IdCurso	Chave estrangeira para **PortalEnsino.Curso**, referencia **Id**; representa o curso associado ao usuário

Tabela 11 – BibliotecaDigital.Usuarios

BibliotecaDigital.Usuarios	
Id	Chave primária; o identificador do usuário
IdUsuarioPortal	Chave estrangeira para **PortalCore.Usuario**, referencia **Id**; representa o usuário associado ao usuário que acessa o sistema
Tipo	Texto; tipo do usuário do sistema

Tabela 12 – PortalEnsino.Curso

PortalEnsino.Curso	
Id	Chave primária; o identificador do curso

Tabela 13 – PortalCore.Pessoa

PortalCore.Pessoa	
Id	Chave primária; o identificador da pessoa

Tabela 14 – PortalCore.Usuario

PortalCore.Usuario	
Id	Chave primária; o identificador do usuário

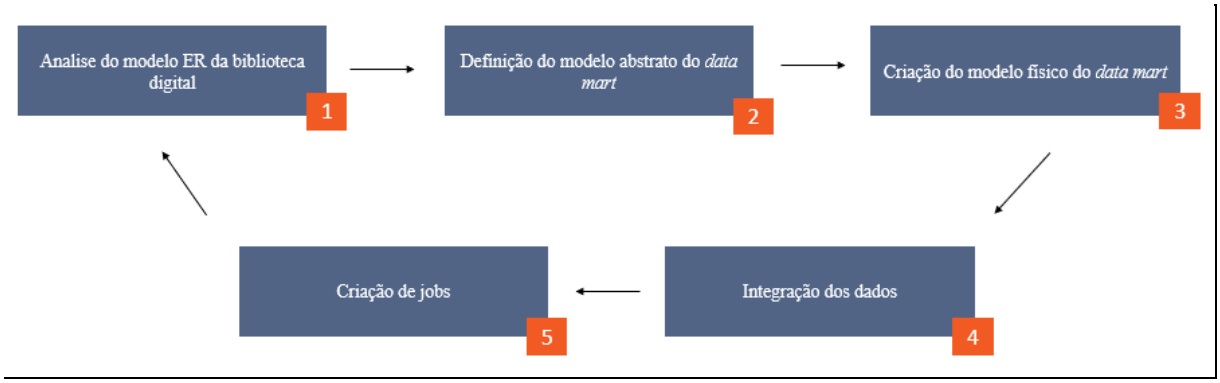
O banco de dados da biblioteca digital conta com 297 documentos (trabalhos de conclusão de curso) cadastrados ao longo de 6 anos, iniciando em 2013, pelos cursos de Biomedicina, Ciência da Computação, Direito, Educação Física, Engenharia Civil, Engenharia de Minas, Engenharia de Software, Farmácia, Odontologia e Sistemas de Informação.

Os bancos de dados apresentados serviram como base para o processo de desenvolvimento do *data mart* e criação do processo ETL.

3.2 Procedimentos

Os procedimentos necessários para o desenvolvimento de um *data mart* e automatização do processo *ETL* foram divididos em cinco etapas que são ilustradas pela Figura 16.

Figura 16. Metodologia.



A Figura 16 apresenta as cinco etapas que foram utilizadas no desenvolvimento do *data mart* tendo como base os dados referentes a base de dados da biblioteca digital do CEULP/ULBRA:

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*, conhecendo os dados armazenados nesse estágio os componentes da estrutura do *data mart* como: Arquitetura do *data mart*; Tipo de implementação; Definição de nível de granularidade de dados; Definição dos fatos; Definição das dimensões, que participam do fato; Definição de medidas/métricas; e Definição do modelo;
2. **definição do modelo abstrato do *data mart***, criado a partir dos elementos básicos da modelagem multidimensional, já mencionados na seção 2.1.4. Nesta etapa foram definidas dimensões, fatos, métricas e modelo estrela para a constituição do *data mart*;

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*.

O ciclo se repetiu à medida que novas análises foram necessárias para a descoberta de novos elementos constituintes ao *data mart*.

Na análise do modelo E-R da Biblioteca Digital realizada para facilitar a identificação de hierarquias e apoiar a modelagem das dimensões do *data mart*. A partir do modelo Entidade-Relacionamento (E-R), foi possível definir:

1. a arquitetura do *data mart*;
2. o tipo de implementação;
3. o cálculo da granularidade de cada tabela fato;
4. as dimensões associadas a cada tabela fato, identificar os atributos que compõem a dimensão através das entidades pertencentes ao modelo, além das cardinalidades entre os atributos.
5. a especificação do fato;
6. identificação de medidas/métricas; e
7. tipo de modelo multidimensional.

Nesta etapa de Definição do Modelo abstrato do *Data Mart* foi realizada a modelagem multidimensional, partindo da identificação dos elementos básicos como: dimensões; fatos; métricas e modelo. O modelo estrela foi definido devido à não existência de normalização, onde a dimensão possui relacionamento de um para muitos fatos.

Os resultados da análise realizada na etapa de definição do modelo abstrato do *data mart* foram implementados na estrutura do *data mart*, utilizando o SQL Server 2014, para armazenamento dos dados posteriores a esta etapa.

O processo de integração dos dados foi realizado utilizando a ferramenta *Pentaho Data Integration – PDI*. Para a integração dos dados o processo foi divido em duas etapas: integração dos dados para criação das dimensões; integração dos dados para criação do fato.

Por fim, o processo de automatização do processo ETL, foi realizado após a criação do fato e das dimensões, utilizando a ferramenta *Pentaho Data Integration – PDI*, com a criação de *jobs* no *kettle*.

4 RESULTADOS E DISCUSSÃO

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. Esta seção apresenta os resultados do desenvolvimento do *data mart* utilizando a ferramenta SQL Server 2014, assim como os resultados do desenvolvimento e automatização do processo *ETL* utilizando a ferramenta *Pentaho Data Integration – PDI/Kettle.* As ferramentas mencionadas possuem funcionalidades que permitiram o desenvolvimento deste trabalho.

4.1 Modelo lógico do *data mart*

A partir das análises dos dados foram levantadas necessidades que direcionaram a criação do modelo lógico:

1. Em quantas publicações de trabalhos do curso de Ciência da Computação o professor X participou como orientador;
2. No 1º semestre do ano de 2017 o curso de Sistemas de Informação teve quantas publicações de trabalhos em que o professor X tenha participado como membro de banca;
3. Qual o percentual publicações de trabalhos realizados com o tema X definido pela palavra-chave Y, do curso de Odontologia;
4. Qual o índice de crescimento de publicações de trabalhos de determinado tema X definido pela palavra-chave Y, do curso Direito desde o ano 2013.

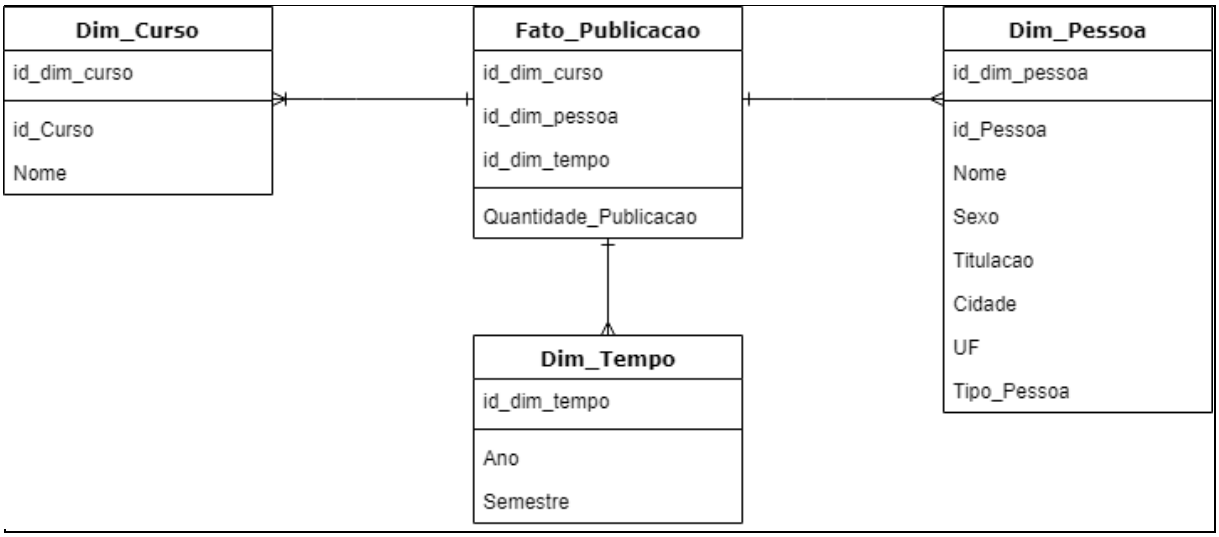
A partir das necessidades levantadas foi possível a identificação dos elementos do modelo multidimensional do *data mart*. As necessidades abordadas apresentam o fato **publicação** em comum. Este é um fato, pois possui valores numéricos como quantidade, percentual e índice para descrevê-lo, e seus valores são mutáveis e evolutivos ao longo do tempo, podendo ser analisado essa evolução ao longo de um espaço de tempo.

As dimensões que participam deste fato **curso**, **pessoa** e **tempo** foram identificadas definindo pontos cardeais do fato, tendo como exemplo a 1º afirmação feita, “Em quantas publicações de trabalhos do curso de Ciência da Computação o professor X participou como orientador”, os respectivos pontos analisados a partir dessa necessidade foram: o fato desta necessidade; a ação da solicitação; a medida que representa o fato; e a evolução da medida ao longo do tempo.

O modelo lógico criado segue o modelo estrela, com o fato publicação no centro e ao seu redor estão dispostas as dimensões curso, pessoa e tempo. À medida que compõe o fato é quantidade de publicação, um valor aditivo que pode ser manipulado algebricamente.

O modelo lógico apresentado na Figura 17, evidencia o fato e as dimensões do *data mart* dispostas no modelo estrela.

Figura 17. Modelo lógico do *data mart*.



O modelo lógico apresentado pela figura indica o fato, suas dimensões e seus campos, apresentados em detalhes a seguir nas tabelas 15 a 18.

Tabela 15 – Fato.Publicacao

Fato_Publicacao	
id_dim_curso	Chave técnica; o identificador da dimensão Dim_Curso
id_dim_pessoa	Chave técnica; o identificador da dimensão Dim_Pessoa
id_dim_tempo	Chave técnica; o identificador da dimensão Dim_Tempo
Quantidade_Publicacao	Campo; contém informação modificada a partir das dimensões

Tabela 16 – Dim.Curso

Dim_Curso	
id_dim_curso	Chave técnica; o identificador da dimensão
id_Curso	Chave; usada para identificar curso em PortalEnsino.Cursos.Id
Nome	Campo; contém informação real de PortalEnsino.Cursos.Nome

Tabela 17 – Dim.Pessoa

Dim_Pessoa	
id_dim_pessoa	Chave técnica; o identificador da dimensão
Id_Pessoa	Chave; usada para identificar pessoa em PortalCore.Pessoa.Id
Nome	Campo; contém informação real de PortalCore.Pessoa.Nome
Sexo	Campo; contém informação real de PortalCore.Pessoa.Sexo
Titulacao	Campo; contém informação real de PortalCore.Pessoa.Titulacao
Cidade	Campo; contém informação real de PortalCore.Pessoa.Cidade
UF	Campo; contém informação real de PortalCore.Pessoa.UF
Tipo_Pessoa	Campo; contém informação real de PortalCore.Pessoa.TipoPessoa

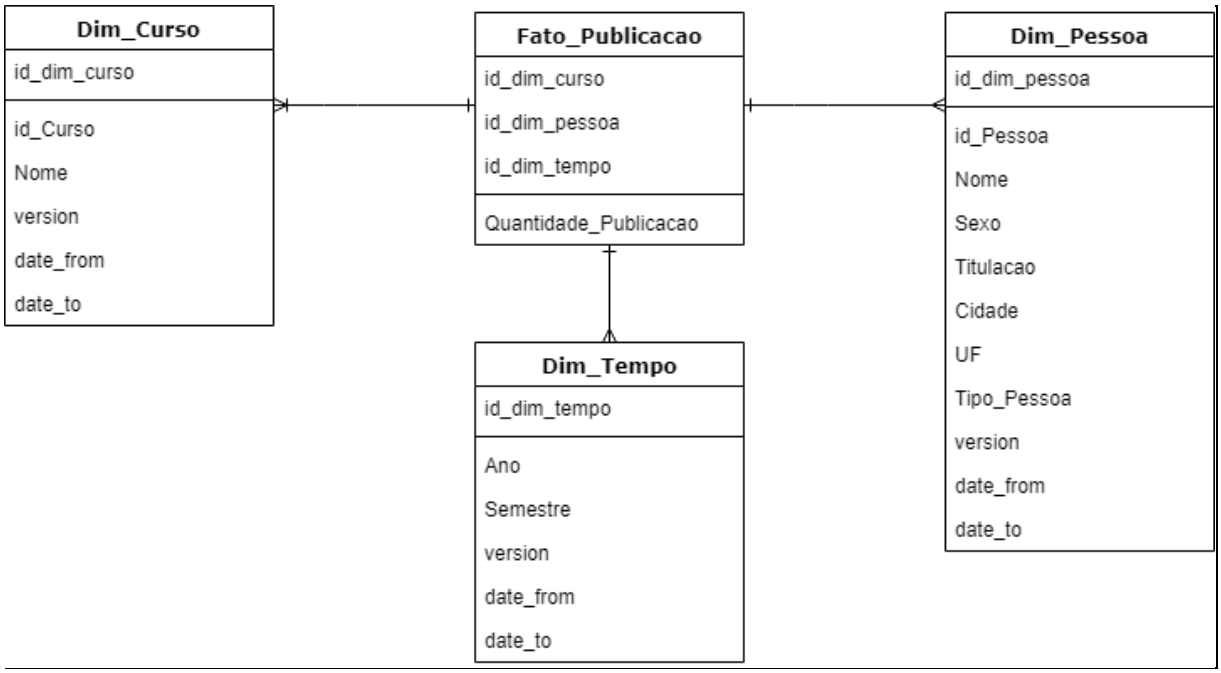
Tabela 18 – Dim.Tempo

Dim_Tempo	
id_dim_tempo	Chave técnica; o identificador da dimensão
Ano	Campo; contém informação real criada a partir do processo ETL
Semestre	Campo; contém informação real criada a partir do processo ETL

4.2 Modelo físico do *data mart*

O modelo lógico foi implementado para a criação do modelo físico do *data mart*. O modelo físico apresentado na Figura 18, evidencia o fato e as dimensões do *data mart* implementadas dispostas no modelo estrela.

Figura 18. Modelo físico do *data mart*.



O modelo físico apresentado pela figura indica o fato, suas dimensões e seus campos, apresentados em detalhes na seção anterior. Nesta implementação as dimensões **Dim_Curso**, **Dim_Pessoa** e **Dim_tempo** passaram a ter novas propriedades para o gerenciamento do processo ETL, os campos são apresentados em detalhes a seguir nas tabelas 19, 20 e 21.

Tabela 19 – Dim.Curso2

Dim_Curso	
version	Campo; versão da entrada da dimensão curso (um número de versão)
date_from	Campo; contém a data de início da validade da carga da dimensão curso
date_to	Campo; contém a data final da validade da carga da dimensão curso

Tabela 20 – Dim.Pessoa2

Dim_Pessoa	
version	Campo; versão da entrada da dimensão pessoa (um número de versão)
date_from	Campo; contém a data de início da validade da carga da dimensão pessoa
date_to	Campo; contém a data final da validade da carga da dimensão pessoa

Tabela 21 – Dim.Tempo2

Dim_Tempo	
version	Campo; versão da entrada da dimensão tempo (um número de versão)
date_from	Campo; contém a data de início da validade da carga da dimensão tempo
date_to	Campo; contém a data final da validade da carga da dimensão tempo

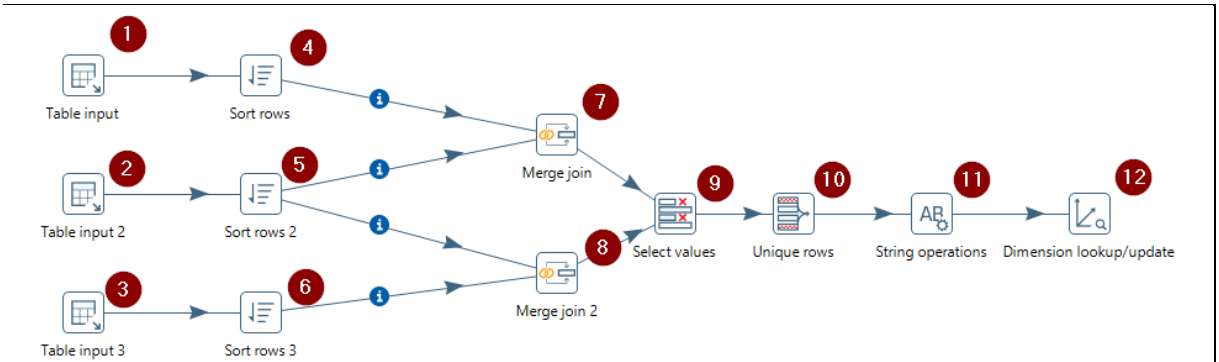
4.3 Processo ETL com *Kettle*

O processo ETL com *kettle* foi a criação do fluxo de trabalho com blocos para integrar os dados, usando a rede de tarefas logicas para criar etapas de leitura, tratamento e carga dos dados.

4.3.1 Processo de integração da Dimensão Pessoa

O processo de integração da dimensão Pessoa é a criação de um processo ETL utilizando o *kettle* para a criação de transformações que integram os dados resultando na dimensão pessoa. A seguir será apresentada na Figura 19, o processo ETL da dimensão Pessoa.

Figura 19. Processo ETL da dimensão Pessoa.



Do processo ETL apresentado na Figura 19, pode-se destacar 12 marcações que representam as etapas do fluxo de trabalho para a integração de dados da dimensão Pessoa, a marcação (1) apresenta a etapa *Table input*, lê informações da tabela MembrosDeBancas do banco de dados BibliotecaDigital, usando instruções *SQL*, conecta-se através de um *hop*, que permite a passagem de metadados de uma etapa para outra, com a etapa de marcação (4) *Sort rows*, ordena as linhas com base no campo IdPessoa em ordem crescente, conecta-se a etapa de marcação (7) *Merge join*, executa uma junção entre conjuntos de dados provenientes das etapas de marcação (4 e 5).

A marcação (2) apresenta a etapa *Table input* 2, lê informações da tabela Pessoa do banco de dados PortalCore, usando instruções *SQL*, conecta-se através de um *hop* com a etapa de marcação (5) *Sort rows* 2, ordena as linhas com base no campo Id em ordem crescente, conecta-se a etapa de marcação (7) *Merge join*, executa uma junção entre conjuntos de dados provenientes das etapas de marcação (4 e 5) e conecta-se a etapa de marcação (8) *Merge join* 2, executa uma junção entre conjuntos de dados provenientes das etapas de marcação (5 e 6).

A marcação (3) apresenta a etapa *Table input* 3, lê informações da tabela AutoresDeDocumentos do banco de dados BibliotecaDigital, usando instruções *SQL*, conecta-se através de um *hop* com a etapa de marcação (6) *Sort rows* 3, ordena as linhas com base no campo IdPessoa em ordem crescente, conecta-se a etapa de marcação (8) *Merge join*, executa uma junção entre conjuntos de dados provenientes das etapas de marcação (5 e 6).

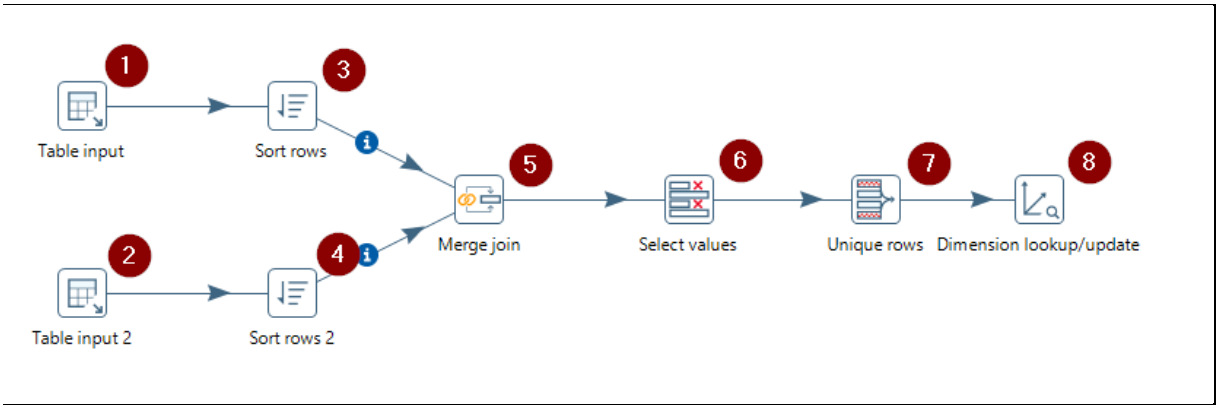
A partir da marcação (9) *Select values*, seleciona-se os valores resultantes da junção das etapas de marcação (7 e 8), executando ações de seleção, remoção e renomeação de valores. A marcação (10) *Unique rows*, remove linhas duplicadas do fluxo de entrada e filtra apenas as linhas únicas com dados de entrada para a etapa. A marcação (11) *String operations*, aplica operação de designação do campo Nome para caracteres maiúsculos. Por fim, a marcação (12)

Dimension lookup/update, insere registros na tabela Dim_Pessoa, correspondentes a dimensão pessoa.

4.3.2 Processo de integração da Dimensão Curso

O processo de integração da dimensão Curso é a criação de um processo ETL utilizando o *kettle* para a criação de transformações que integram os dados resultando na dimensão curso. A seguir será apresentada na Figura 20, o processo ETL da dimensão Curso.

Figura 20. Processo ETL da dimensão Curso.



Do processo ETL apresentado na Figura 20, pode-se destacar 8 marcações que representam as etapas do fluxo de trabalho para a integração de dados da dimensão Curso, a marcação (1) apresenta a etapa *Table input*, lê informações da tabela Curso do banco de dados PortalEnsino, usando instruções *SQL*, conecta-se através de um *hop*, que permite a passagem de metadados de uma etapa para outra, com a etapa de marcação (3) *Sort rows*, ordena as linhas com base no campo Id em ordem crescente, conecta-se a etapa de marcação (5) *Merge join*, executa uma junção entre conjuntos de dados provenientes das etapas de marcação (3 e 4).

A marcação (2) apresenta a etapa *Table input* 2, lê informações da tabela Documentos do banco de dados BibliotecaDigital, usando instruções *SQL*, conecta-se através de um *hop* com a etapa de marcação (4) *Sort rows* 2, ordena as linhas com base no campo IdCurso em ordem crescente, conecta-se a etapa de marcação (5) *Merge join*, executa uma junção entre conjuntos de dados provenientes das etapas de marcação (3 e 4).

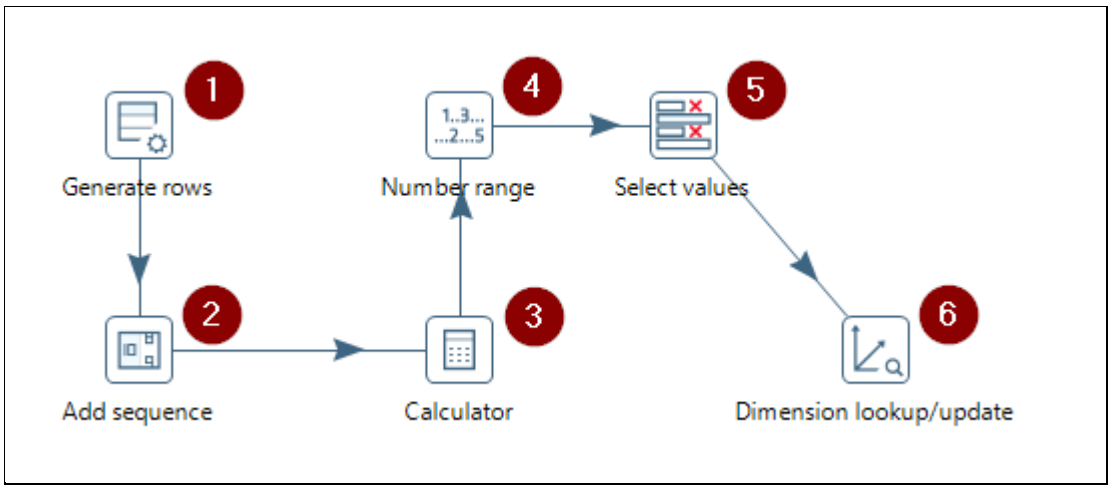
A partir da marcação (6) *Select values*, seleciona-se os valores resultantes da junção da etapas de marcação (5), executando ações de seleção de valores. A marcação (7) *Unique rows*, remove linhas duplicadas do fluxo de entrada e filtra apenas as linhas únicas com dados de

entrada para a etapa. Por fim, a marcação (8) *Dimension lookup/update*, insere registros na tabela Dim_Curso, correspondentes a dimensão curso.

4.3.3 Processo de integração da Dimensão Tempo

O processo de integração da dimensão Tempo é a criação de um processo ETL utilizando o *kettle* para a criação de transformações que resultem na dimensão tempo. A seguir será apresentada na Figura 21, o processo ETL da dimensão Tempo.

Figura 21. Processo ETL da dimensão Tempo.

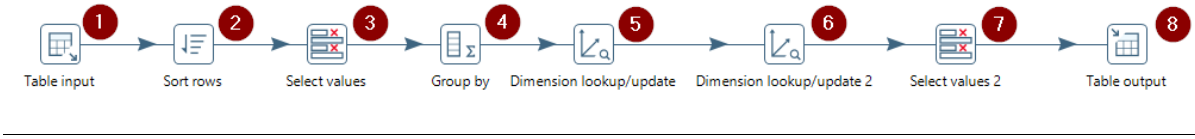


Do processo ETL apresentado na Figura 20, pode-se destacar 6 marcações que representam as etapas do fluxo de trabalho para a integração de dados da dimensão Tempo, a marcação (1) *Generate rows*, gera saídas de linhas, usada para gerar 3287 linhas, conecta-se através de um *hop* com a etapa de marcação (2) *Add sequence*, adiciona uma sequência ao fluxo, conecta-se com a etapa de marcação (3), *Calculator*, que executa nos valores dos campos das linhas geradas funções pré-definidas para calcular uma data, dia do mês, dia do ano, mês, ano e semana para na etapa de marcação (4) *Number ranger*, atribua intervalos com base nos meses calculados para se obter o campo Semestre, assim na etapa de marcação (5) *Select values*, seleciona-se os valores resultantes das etapas de marcação (3 e 4). Por fim, a marcação (6) *Dimension lookup/update*, insere registros na tabela Dim_Tempo, correspondentes a dimensão tempo.

4.3.4 Processo de integração do Fato Publicação

O processo de integração do fato Publicação é a criação de um processo ETL utilizando o *kettle* para integrar os dados resultando das dimensões pessoa, curso e tempo. A seguir será apresentada na Figura 22, o processo ETL do fato Publicação.

Figura 22. Processo ETL do fato Publicação.

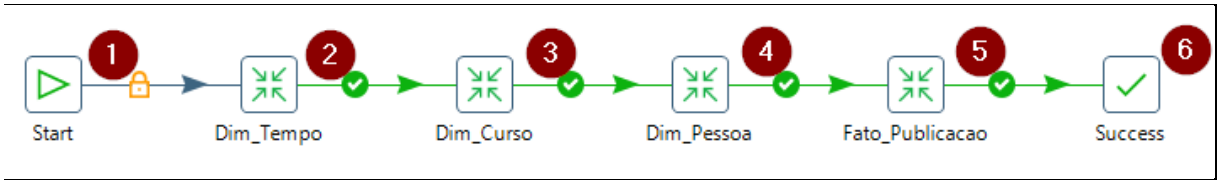


Do processo ETL apresentado na Figura 22, pode-se destacar 8 marcações que representam as etapas do fluxo de trabalho para a integração de dados do fato Publicação, a marcação (1) apresenta a etapa *Table input*, lê informações da dimensão Pessoa do *data mart*, usando instruções *SQL*, conecta-se através de um *hop* com a etapa de marcação (2) *Sort rows*, ordena as linhas com base no campo IdPessoa em ordem crescente, conecta-se a etapa de marcação (3) *Select values*, seleciona-se os valores desejados da etapa de marcação (1), assim na etapa de marcação (4) *Group by*, agrupa-se linhas com base no campo IdPessoa para gerar valores agregados para o grupo, contato o numero de publicações de documentos, assim o resultado desta etapa e conectado através do *hop* para a etapa de marcação (5) *Dimension lookup/update*, usada para pesquisar a chave técnica id_dim_curso da dimensão Curso, assim o resultado desta etapa e conectado através do *hop* para a etapa de marcação (6) *Dimension lookup/update* 2, usada para pesquisar a chave técnica id_dim_tempo da dimensão Tempo, na etapa de marcação (7) *Select values* 2, seleciona-se os valores para a inserção de registros na etapa de marcação (8) *Table output*, insere registros na tabela Fato_Publicacao, correspondentes ao fato publicação.

4.3.5 Automatização dos processos de integração com *jobs*

A automatização dos processos de integração com jobs é a criação de um agendamento utilizando o *kettle* para coordenar execuções e recursos de atividades ETL. A seguir será apresentada na Figura 23, a automatização dos processos de integração com *job*.

Figura 23. Automatização dos processos de integração com *job*.



Da automatização dos processos de integração com *job* apresentado na Figura 23, pode-se destacar 6 marcações que representam as etapas do fluxo de trabalho para a integração dos processos ETL das dimensões de Tempo, Curso, Pessoa e do fato Publicação, a marcação (1) *Start*, é o ponto de partida para a execução do *job*, na etapa de marcação (2) Dim_Tempo, executa a transformação previamente definida, esta etapa é o ponto de acesso para a atividade ETL da dimensão Tempo, na etapa de marcação (3) Dim_Curso, executa a transformação previamente definida, esta etapa é o ponto de acesso para a atividade ETL da dimensão Curso, na etapa de marcação (4) Dim_Pessoa, executa a transformação previamente definida, esta etapa é o ponto de acesso para a atividade ETL da dimensão Pessoa, na etapa de marcação (5) Fato_Publicacao, executa a transformação previamente definida, esta etapa é o ponto de acesso para a atividade ETL do fato Publicação, por fim, na etapa de marcação (6) *Success*, limpa qualquer estado de erro encontrado em um trabalho e o força para um estado de sucesso.

5 CONSIDERAÇÕES FINAIS

O presente trabalho teve como resultado o desenvolvimento de um *data mart* e automatização do processo *ETL* para centralizar os dados referentes à produção acadêmica disponíveis na base de dados da Biblioteca Digital do CEULP/ULBRA.

O trabalho foi conduzido em fases com início no entendimento do contexto, problema do trabalho, objetivos definidos e dos procedimentos criados para alcançar os resultados, os materiais utilizados vieram a ser de muita importância com destaque para o *pentaho/kettle*, possuindo uma documentação bem estruturada de cada etapa e transformação da ferramenta, para a criação do processo ETL, além de possuir uma interface gráfica com a possibilidade de projetar uma rede de fluxos de dados sem a necessidade da utilização de programação via código.

Em virtude aos objetivos de modelar e implementar o *data mart*, e automatizar o processo ETL, eram desafios que a medida do tempo foram superados a cada etapa do procedimento, com a analise do modelo E-R da Biblioteca Digital foi possível modelar o modelo logico para o *data mart*, assim foi possível a criação do modelo fisico. Com a estrutura implementada houve inicio aos processos de integração e por fim a automatização desses processos através da criação de *jobs*. O que se iniciou através da teoria e entendimento do referencial teórico foi aplicado e desenvolvido a fim de obter os resultados determinados no início do trabalho.

Durante a elaboração deste trabalho foram identificados alguns pontos que poderiam ser explorados para que houvesse um aperfeiçoamento nos resultados obtidos. Por este trabalho se tratar do desenvolvimento de um *data mart* para um setor específico dentro da instituição do CEULP/ULBRA, já foi pensado para integrar a outros, pois sua arquitetura é do tipo integrada que da suporte a conexão de outros *data marts*. Como a Biblioteca Digital outros sistemas podem ser estudados para a criação de *data marts*, podendo assim instituírem um só *data warehouse*.

Como o foco deste trabalho foi baseado na estrutura do *data mart*, é recomendado que haja um trabalho com utilização de ferramentas relacionadas as consultas para o usuario final, como elaborações de relatórios, pesquisas informativas, analises de desempenho, mineração de dados e visualização de dados com base nos dados armazenados neste *data mart*.

REFERÊNCIAS

ABREU, Fábio Silva Gomes da Gama e. Desmistificando o conceito de etl. **Revista de Sistemas de informação,** n. 02, p. 6, jul./dez. 2008. Disponível em: <http://www.fsma.edu.br/si/Artigos/V2_Artigo1.pdf>. Acesso em: 02 abr. 2017.

DIAS, Sandra Aparecida. **Integração Semântica de Dados Através de Federação de Ontologias.** 2006, p. 79. Dissertação (Mestrado pelo programa de Pós-Graduação em informática) - PUC RIO, Rio de Janeiro, 2006. Disponível em: <http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0321535_06_pretextual.pdf>. Acesso em: 05 de maio 2017.

ELEUTERIO, Marco Antonio Masoller. **Sistemas de Informações gerenciais na atualidade.** Curitiba: InterSaberes, 2015.

ELMASRI, Ramez; NAVATHE, Shamkant 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***. INForum 2010 – II Simpósio de Informática**, p. 757-765, set. 2010. Disponível em: <http://inforum.org.pt/INForum2010/papers/sistemas-inteligentes/Paper080.pdf>. Acesso em: 02 abr. 2017.

FREITAS, Gilmar Meira. **Uma ferramenta de apoio à modelagem de dados dimensional. 2001**. Dissertação (Mestrado em Administração Pública) – Fundação João Pinheiro, Belo Horizonte, 2002.

GONÇALVES, Marcio. **Extração de Dados Para Data Warehouse.** Palmas: Axcel Books do Brasil p. 147, 2003.

INMON, W. H. The data warehouse and data mining. **Communications of the ACM**, v. 39, n.11, p.49-50, 1996.

INMON, W. H.; HACKATHORN, R. D. **Como usar o Data Warehouse**. IBPI, 1997.

INMON, W. H. **Como construir o data warehouse**. Indianapolis: Wiley Publishing Inc., 1997.

KLEON, Austin. **Roube como um artista**: 10 dicas sobre criatividade. Rio de Janeiro: Rocco, 2013.

KIMBALL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit.** Rio de Janeiro: Editora Campus, ed. 2, p. 494, 2002.

KIMBALL, Ralph; CASERTA, Joe. **The Data Warehouse ETL Toolkit:** Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Indianapolis: Wiley Publishing, Inc., p. 467, 2004.

KIMBALL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit:** The Definitive Guide to Dimensional Modeling. Indianapolis: John Wiley & Sons, Inc, p. 543, 2013.

MACHADO, Felipe Nery Rodrigues. **PROJETO DE DATA WAREHOUSE:** Uma visão Multidimensional. São Paulo: Editora Érica Ltda., p. 248, 2000.

MACHADO, Felipe Nery Rodrigues. **Tecnologia e Projeto de Data Warehouse:** Uma visão multidimensional. São Paulo: Editora Érica Ltda, 2004.

MARTINS, Bárbara. **Tomada de decisão: analisando o uso de sistemas de informação na empresa joagro ferragens de Estrela/RS.** 2014. p. 75. Monografia (Bacharel em Administração) - Centro Universitário UNIVATES, Lajeado, 2014. Disponível em: <https://www.univates.br/bdu/bitstream/10737/786/1/2014BarbaraMartins.pdf>. Acesso em: 02 abr. 2017.

PENTAHO, **Open Source Business Intelligence**. Disponível em: http://kettle.pentaho.com/ Acesso em: 29 mar. 2019.

RUSSOM, Philip. **Unifying the Practices of Data Profiling, Integration, and Quality (dPIQ)**. out. 2007. Disponível em: <http://download.101com.com/pub/tdwi/Files/TDWI_Monograph_DataFlux_Oct2007.pdf>. Acesso em: 28 fev. 2017.

RUSSOM, Philip. **Complex Data: A New Challenge for Data Integration**. nov. 2007. Disponível em: <http://download.101com.com/pub/TDWI/Files/TDWI_Monograph_ComplexData_November2007.pdf>. Acesso em: 28 fev. 2017.

SINGH, Harry S. **Data Warehouse:** conceitos, tecnologias, implementação e gerenciamento. São Paulo: Makron Books, p. 367, 2001.