CENTRO PAULA SOUZA

FACULDADE DE TECNOLOGIA DE AMERICANA

Curso de Bacharel em Análise de Sistemas e Tecnologia da Informação

Rodrigo Reissler Rocha

MELHORES MÉTODOS PARA CONSTRUÇÃO DE DATA WAREHOUSE

AMERICANA/SP

2014

CENTRO PAULA SOUZA

FACULDADE DE TECNOLOGIA DE AMERICANA

Curso de Bacharel em Análise de Sistemas e Tecnologia da Informação

Rodrigo Reissler Rocha

MELHORES MÉTODOS PARA CONSTRUÇÃO DE DATA WAREHOUSE

Trabalho de Conclusão de Curso desenvolvido em cumprimento à exigência curricular do Curso de Bacharel em Análise de Sistemas e Tecnologia da Informação, sob a orientação da Profa. Dra. Maria Cristina Aranda Batocchio.

Área de concentração: Banco de Dados.

Americana, S. P.

2014

Rocha, Rodrigo Reissler
R56m Melhores métodos para construção de uma data warehouse. / Rodrigo Reissler Rocha. – Americana: 2014.
56f.

Monografia (Graduação de Tecnologia em Análise de Sistemas e Tecnologia da Informação). - - Faculdade de Tecnologia de Americana – Centro Estadual de Educação Tecnológica Paula Souza.
Orientador: Prof. Dr. Maria CristrinaAranda Batocchio

1.Banco de dados I. Batocchio, Maria Cristina ArandaII. Centro Estadual de Educação Tecnológica Paula Souza – Faculdade de Tecnologia de Americana.

CDU: 681.3.07

Rodrigo Reissler Rocha

MELHORES MÉTODOS PARA CONSTRUÇÃO DE DATA WAREHOUSE

Trabalho de graduação apresentado como exigência parcial para obtenção do título de Bacharel em Análise de Sistemas e Tecnologia da Informação pelo CEETEPS/Faculdade de Tecnologia – FATEC/ Americana.
Área de concentração: Banco de Dados.

Americana, 26 de Junho de 2014.

Banca Examinadora:

_____
Profa. Maria Cristina Aranda Batocchio (Presidente)
Doutora
Fatec Americana

_____
Prof. Rodrigo Viviani (Membro)
Graduação
Fatec Americana

_____
Profa. Maria Elizete Luz Saes (Membro)
Meste
Fatec americana

AGRADECIMENTOS

Os meus agradecimentos vão primeiramente a minha orientadora Maria Cristina Aranda Batocchio, pela presteza que demonstrou durante o projeto, orientando para que este Trabalho de Conclusão de Curso fosse finalizado com sucesso.

Em especial aos meus amigos Acácio, Ailton, Jorge Lucas e Rafael pela ajuda e força na troca de informações e pela alegria, demonstrando muita amizade e solidariedade. À minha namorada Luciane, pela dedicação em orientar no desenvolvimento e compreensão de minha ausência devido ao tempo destinado para elaboração deste trabalho.

Aos meus amigos de trabalho que forneceram informações e conhecimento para elaboração deste trabalho.

Por fim, agradeço a Deus pelas oportunidades que foram dadas para ganhar e compartilhar experiências vividas.

DEDICATÓRIA

Dedico este trabalho aos meus pais, amigos e professores, que contribuíram na minha formação acadêmica, profissional e pessoal.

.

RESUMO

Decisões geram consequências, positivas ou negativas. Para diminuir a possibilidade de erros em decisões, pessoas ou empresas procuram analisar fatos relacionados ao assunto em questão. No ambiente corporativo este comportamento torna-se cada vez mais recorrente e o volume de dados utilizados nas análises cresce constantemente, por este motivo que o conceito de Data Warehouse foi criado, de forma a permitir que essa atividade seja desenvolvida indiferente da quantidade ou origem dos dados. A eficiência deste método é garantida pelas diversas etapas de construção, na escolha da arquitetura e do modelo dimensional, que definem a forma de como os dados serão organizados dentro do bando de dados, além do processo de extração, transformação e carga, etapa que pode ser realizada com o auxílio de ferramentas especializadas, garantindo a confiabilidade dos dados e impedindo que os resultados das análises sejam comprometidos por falta de registros ou pela incompatibilidade dos dados obtidos em origens distintas.

Palavras Chave: Banco de Dados, Data Warehouse, *ETL* (Extração Transformação Carga).

ABSTRACT

Decisions generate consequences, positive or negative, to reduce the possibility of errors in decisions, peoples or companies search to analyze facts related to the point in question. In the corporate environment, this activity becomes increasingly recurrent and the volume of data used in the analyzes increases constantly, for this reason that the concept of Data Warehouse was created in order to allow this activity is developed indifferent of the quantity or source of the data . The efficiency of this method is guaranteed by the various stages of construction, the choice of architecture and dimensional model, which define how the data are organized in the data base, beyond extraction, transformation and load process, this step can be execute with the help of specialized tools and should ensure the reliability of data preventing the analysis results are compromised by absence of records or by incompatibility of data obtained disparate sources.

Keywords: Data Base, Data Warehouse, ETL (Extract Transform Load).

Sumário

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BI	Business Intelligence
DM	Data Mart
DW	Data Warehouse
ETL	Extract Transform Load
FTP	File Transfer Protocol
OLAP	On-Line Analytical Processing
OLTP	On-Line Transaction Processing
OWB	Oracle Warehouse Builder
PDI	Pentaho Data Integration
SI	Sistema de Informação
SGBD	Sistema de Gerenciamento de Banco de Dados
SSIS	SQL Server Integration Services
TDI	Talend Data Integration
TI	Tecnologia da Informação
XML	eXtensible Markup Language

Introdução

Desde os primórdios as decisões baseadas em acontecimentos fazem parte do cotidiano dos seres humanos. No ambiente corporativo não é diferente, as empresas estão cientes da influência que as análises de informações podem causar em seus resultados e possuir funcionários qualificados. Uma área de tecnologia de informação bem estruturada, com incentivos à inovação, pode acarretar o aumento de valor à empresa (KRAEMER, 2003).

A tomada de decisão é uma atividade exercida pela gerência dentro de uma organização. Indiferente do contexto, esta ação deve ser baseada em um processo sistematizado, que envolva a análise do problema mediante a obtenção de dados, produção de informação, desenvolvimento de propostas de soluções, estudo de viabilidade, implementação da decisão e levantamento dos resultados obtidos (CHOO, 1998).

Conforme Primak (2008), todo este processo é definido como Business Intelligence (BI):

> Processo inteligente de coleta, organização, análise, compartilhamento e monitoração de dados contidos em Data Warehouse / Data Mart gerando informações para o suporte à tomada de decisões no ambiente de negócios.

Devido ao aumento no volume de dados armazenados por uma corporação e da necessidade de gerar conhecimento através do dados, o conceito de Data Warehouse (DW) surgiu para organizar e facilitar a leitura dos dados. De acordo com Inmon (1997), “um Data Warehouse é um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo, de apoio às decisões gerenciais”.

Os dados contidos em um Data Warehouse (DW) são cópias dos dados transacionais, gerados nos diversos sistemas de apoio durante o processo operacional de uma empresa, estruturados para alimentar relatórios gerenciais. (KIMBALL, 2002).

Os sistemas de origem possuem os dados que serão trabalhados, o Data Warehouse é a estrutura onde os dados já trabalhados são armazenados e o ETL

(Extração, Transformação e Carga de Dados) é o processo que contempla as atividades para estruturação do DW (COREY, 2001).

O objetivo principal deste estudo é pesquisar os melhores métodos para construção de um Data Warehouse, considerando o grande volume de dados que são produzidos e armazenados em diferentes bancos de dados nas grandes empresas, sendo objetivos secundários o levantamento de técnicas de extração, transformação e carga de dados, além de realizar um estudo de caso que contemple a análise de uma ferramenta que ofereça suporte a este processo.

Uma análise minuciosa em artigos, livros e demais publicações sobre este assunto se fez necessário e os resultados estão dispostos em três capítulos. No primeiro são apresentados os conceitos, técnicas e etapas da construção de um Data Warehouse, o segundo capítulo é composto pelo estudo de caso da ferramenta Kettle (Pentaho Data Integration) e o último possui as considerações finais da pesquisa.

1 Data Warehouse

Os primeiros Data Warehouse surgiram na década de 90, com objetivo de realizar tarefas que não eram possíveis no banco de dados de sistemas em operação e para compor uma ferramenta exclusiva de obtenção de dados estratégicos de forma flexível, eficaz e eficiente (SINGH, 2001).

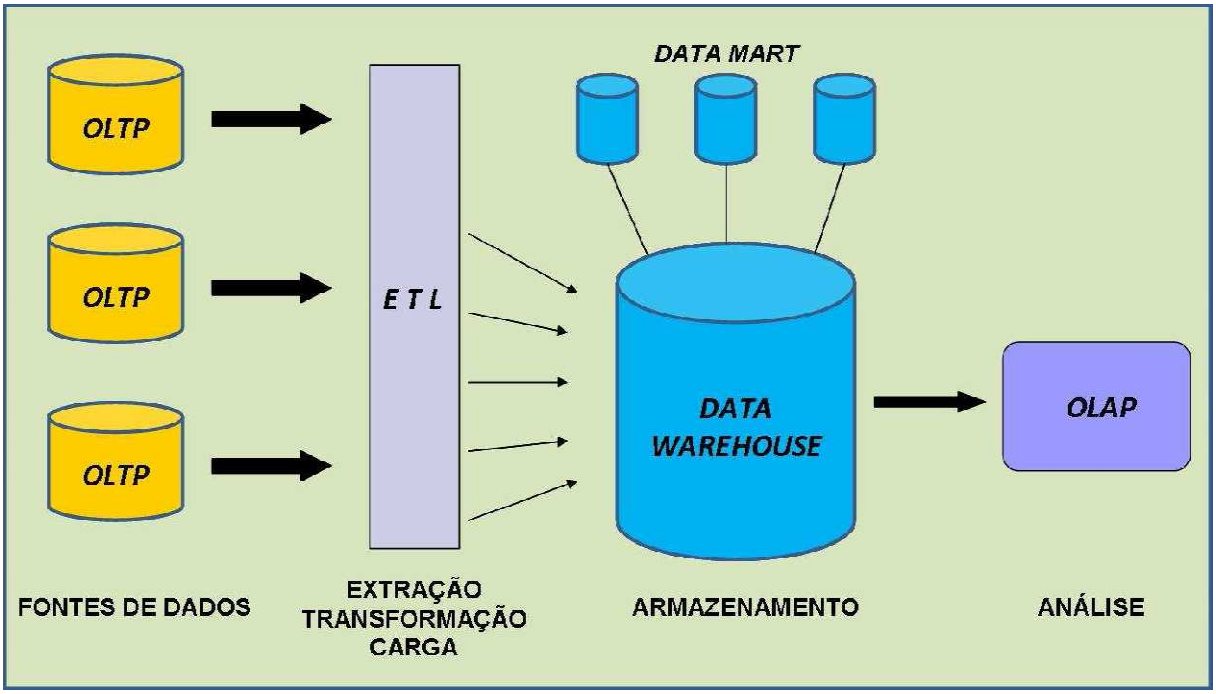
O baixo custo de armazenamento de dados em disco, proporcionou às organizações a viabilidade de guardarem os seus dados por um tempo maior. O grande volume de dados dificulta a análise e demanda maior tempo para transformá-los em informação, isto porque, os dados normalmente estão dispersos, não possuem integração com as demais bases de dados e inexiste uma estrutura para visão integrada. Esses pontos tornaram-se grandes desafios para as empresas e por este motivo a tecnologia de Data Warehouse foi elaborada (GONÇALVES, 2003).

Os dados podem ser originados em diversos sistemas internos ou externos, de forma que o Data Warehouse torne-se um banco de dados especializado, que integra, filtra, gerencia e disponibiliza os dados em um banco de dados paralelo aos sistemas operacionais da empresa (OLIVEIRA, 2002).

O DW tem grande importância em empresas que precisam de respostas rápidas e consistentes auxiliando todo o gerenciamento, sendo essa tecnologia uma das partes fundamentais do conjunto de sistemas de apoio à tomada de decisão, denominado de Business Intelligence (BI) (INMON, 1997).

A Figura 1 demonstra um processo de Business Intelligence, onde a fonte de dados representa os diversos sistemas de processamento de transações em tempo real, *On-Line Transaction Processing* (OLTP), que através do processo de ETL são integrados no DW e estratificados nos Data Marts, dependendo da estratégia adotada, de maneira que ao término do processo são utilizadas as ferramentas para análise dos dados, *On Line Analytical Processing* (OLAP) (BARBIERI, 2001).

Figura 1: Estrutura do Business Intelligence



Fonte: BARBIERI (2001).

De acordo com Corey (2001), um Data Warehouse possui os seguintes aspectos:

- **Integrado:** armazena e integra os dados de diversas fontes em um único ambiente. A integração pode ser complexa devido a possível existência de dados não padronizados, proporcionados pela utilização de sistemas independentes, sem relacionamentos.
- **Não volátil:** garante a permanência da informação em relação ao tempo, impede que usuários realizem qualquer alteração dos dados, de forma que possuam apenas permissão leitura.
- **Orientado ao assunto:** define a disposição dos dados de acordo com os diversos assuntos que existam no contexto da empresa, por exemplo, compras, produção, recursos humanos, vendas, etc.
- **Variável ao tempo:** permite a comparação dos dados em relação ao tempo, assim é possível investigar se uma determinada decisão teve o resultado esperado, analisar se determinada variável pode vir ou não

influenciar no negócio e, a partir da maturidade das análises verificar-se possíveis tendências.

A documentação de toda estrutura e atividades desenvolvidas no ambiente do DW são chamadas de Metadados, pode-se considerar que este conjunto de informações forma um catálogo do Data Warehouse (GONÇALVES, 2003).

Os Metadados são dados de nível mais alto, gerados a partir dos dados de baixo nível existentes na estrutura do Data Warehouse. Eles estão dispostos em diversos formatos para facilitar a compreensão dos vários tipos de usuários.

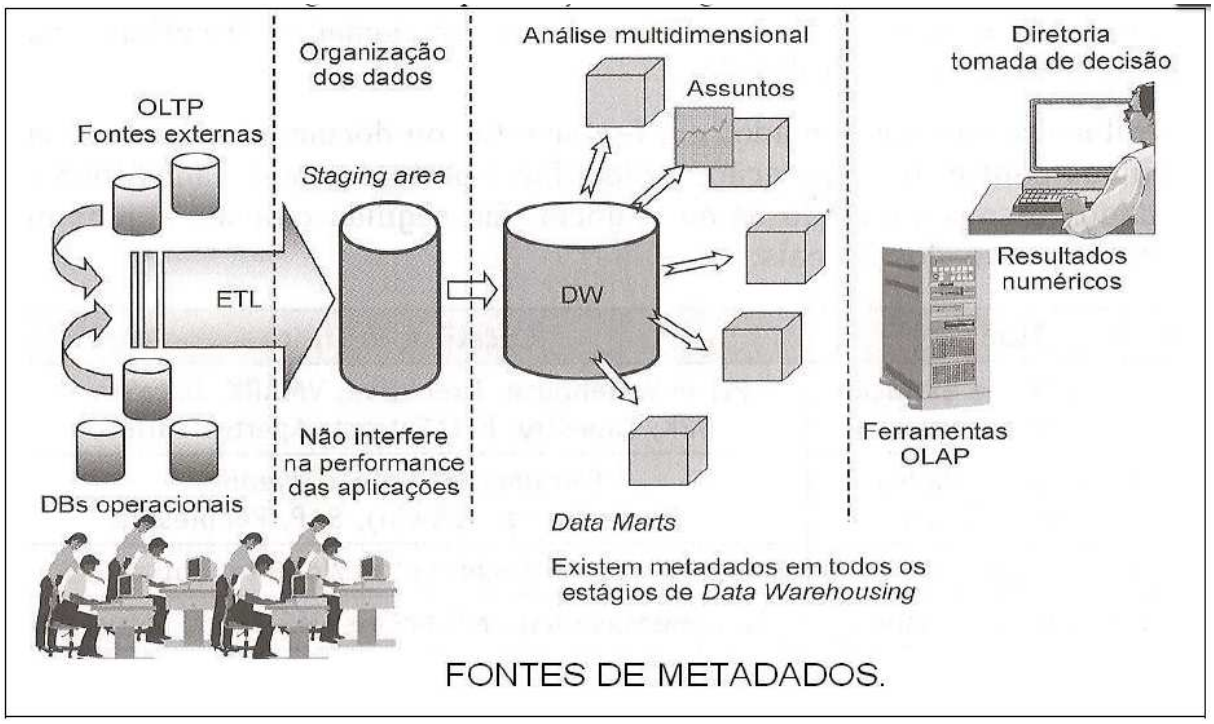
Existem dois tipos de Metadados, os técnicos, que atendem as necessidades de profissionais de tecnologia da informação e os de negócio, para dar suporte aos analistas e administradores de negócio.

Os Metadados técnicos, possuem informações relacionadas à origem dos dados, regras de relacionamento definidas nas entrevistas com usuários, regras dos sistemas utilizados em todo o processo e demais informações que auxiliam os profissionais de TI na manutenção e desenvolvimento do Data Warehouse.

Os Metadados de negócio são dados sobre as padronizações realizadas durante o processo de transformação, disposição dos dados no DW, origem dos dados, entre outras informações necessárias para compreensão dos dados existentes no Data Warehouse, com objetivo de dar suporte aos usuários que farão a análise dos dados (MACHADO, 2008).

A Figura 2 representa a origem dos Metadados, em todos os estágios da construção do DW.

Figura 2: Origem dos Metadados



Fonte: Machado (2008).

Observando a Figura 2 é possível identificar várias fontes de metadados, como exemplo, os dados sobre as transformações realizadas na etapa de ETL, os relacionamentos dos dados no DW na etapa da organização dos dados, os dados da arquitetura entre os DW e os Data Mart na etapa da análise multidimensional e os dados sobre as considerações de negócio utilizada nos relatórios na etapa da diretoria de tomada de decisões.

Após o levantamento realizado sobre as características de um DW e da forma como a documentação é construída, o próximo tópico aborda a estruturação do Data Warehouse que tem como objetivo organizar os dados de maneira ampla facilitando o gerenciamento de acessos.

1.1 Arquitetura

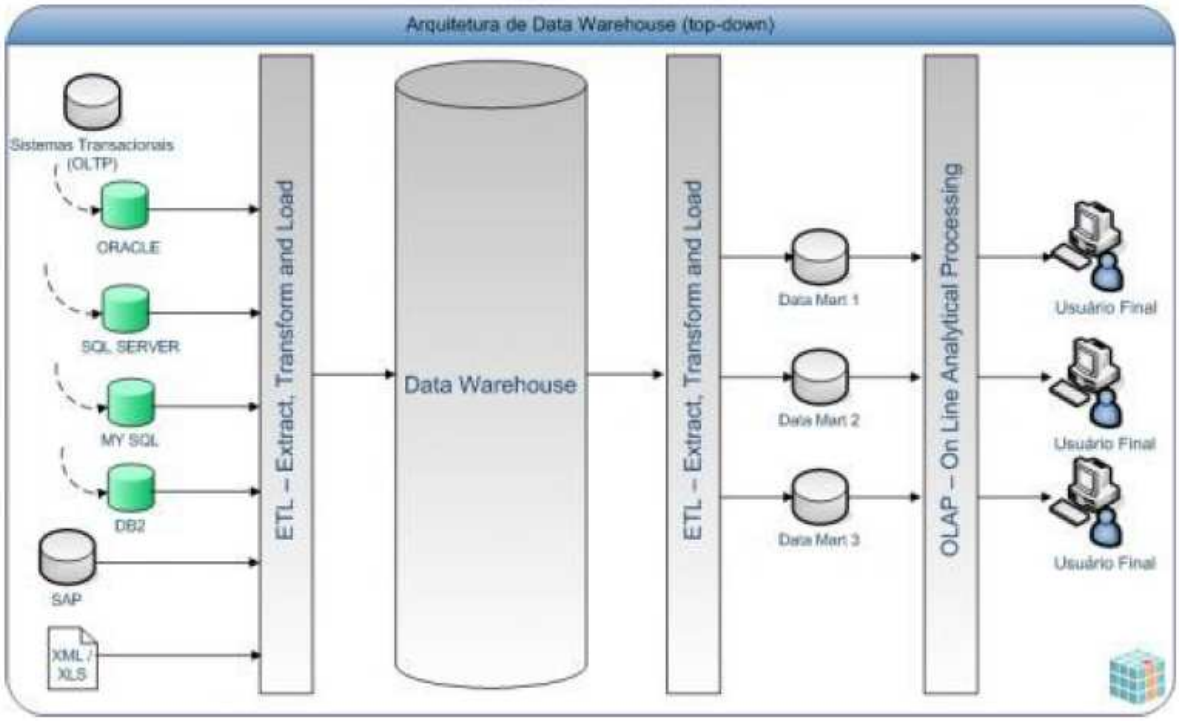
Os dados de um DW são organizados por assunto e para cada assunto pode existir um Data Mart (DM), portanto, o DM possui parte do conteúdo de um Data Warehouse, representando uma determinada área da empresa e por possuir dados de uma parte do negócio, possibilitando o acesso às informações de forma seletiva. Mediante estas características, um DW é formado com dados de vários Data Marts, onde a quantidade é determinada pelo número de áreas consideradas estratégicas para a empresa (HAMMERGREN, 2009).

Howson (2008) alega que a construção do Data Mart pode ocorrer com a utilização dos dados de um Data Warehouse já implementado ou o Data Mart pode popular um Data Warehouse

Bonomo (2009) elenca dois tipos de arquitetura de DW, a Top-Down e a Bottom-Up.

- **Top-Down:** é quando um Data Warehouse é criado e depois segmentado, de maneira que a sua divisão crie bancos menores orientados por assuntos aos departamentos. A Figura 3 demonstra o processo da arquitetura Top-Down, onde os dados podem ser extraídos de um ERP, de arquivos locais (TXT, XML) ou sistemas transacionais (OLTP). Estes dados são filtrados, padronizados e carregados no DW através do processo de extração, transformação e carga de dados (ETL). Após concretizar todas estas etapas os Data Mart são criados a partir dos dados contidos no DW (BONOMO, 2009).

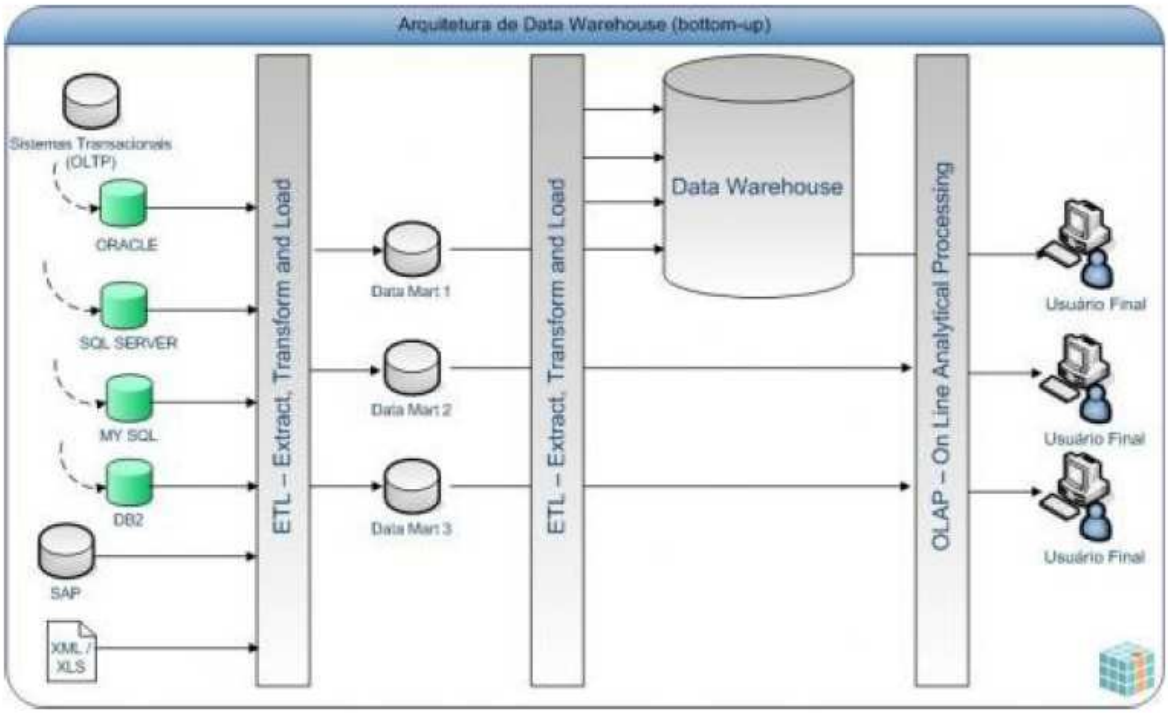
Figura 3: Arquitetura Top-Down.



Fonte: Bonomo (2009).

- **Bottom-Up:** conforme exemplificado na figura 4, nessa arquitetura o processo ETL alimenta inicialmente cada Data Mart que são consolidados em um único banco de dados, através de outro processo ETL. Os Data Mart e Data Warehouse criados a partir deste conceito são utilizados como fonte de dados pelas ferramentas OLAP, sistemas de interação com o usuário que possuem funcionalidades específicas para análise e apresentação de dados (BONOMO, 2009).

Figura 4: Arquitetura Bottom-Up



Fonte: Bonomo (2009).

Com os conceitos de arquitetura esclarecidos o próximo passo é compreender com mais detalhe a disposição dos dados no DW, portanto, o próximo item descreve sobre a modelagem dimensional.

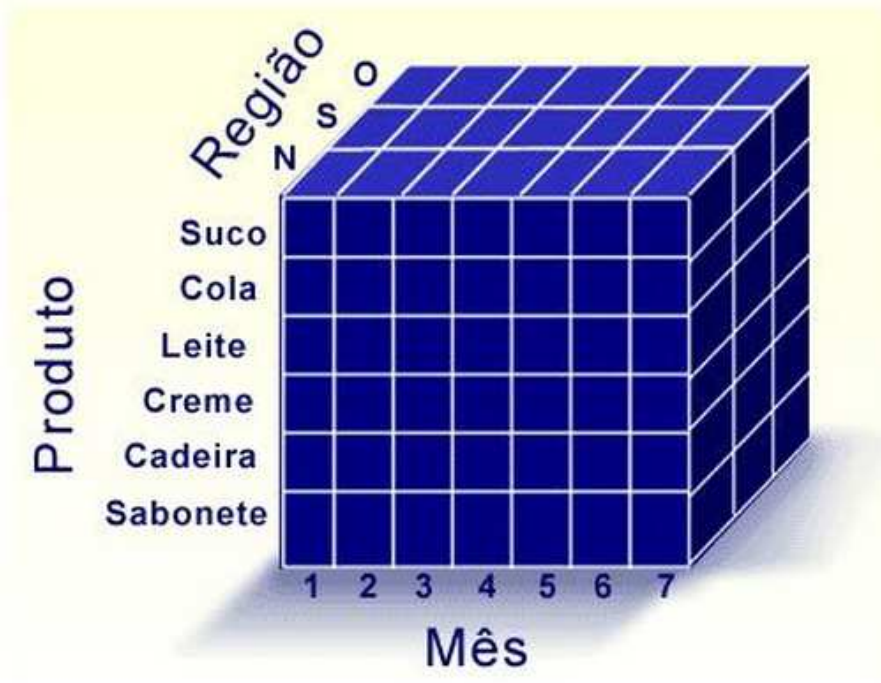
1.2 Modelagem Dimensional

A Modelagem Dimensional é uma técnica antiga que torna os bancos de dados compreensíveis, além disso, permite o desenvolvimento de modelos lógicos de dados para melhorar o desempenho de consultas. Esta metodologia é utilizada desde a década de 1970 e foi elaborada para atender a necessidade humana, facilitando a interpretação dos profissionais de Tecnologia da Informação, consultores, fornecedores e usuários finais (KIMBALL, 2002).

Um modelo dimensional possui uma tabela principal, chamada de tabela de fatos, que se relaciona com outras tabelas. As tabelas secundárias, nomeadas como tabelas de dimensões que possuem apenas uma junção que se conecta com a tabela de fatos (KIMBALL, 2002).

Com a aplicação do modelo, os dados podem ser visualizados no formato de um cubo, onde as dimensões do cubo representam um determinado contexto que envolve os fatos, e os relacionamentos entre as dimensões representando as medidas do fato. Deve-se aplicar a quantidade de dimensões necessárias para se representar um fato. A Figura 5 apresenta um exemplo com as dimensões de localização, produto e tempo, divididas por níveis, assim permite a elaboração de análises analíticas ou gerenciais (NARDI, 2000).

Figura 5: Representação de um cubo.



Fonte: Nardi (2000).

Os elementos essenciais de um modelo dimensional são a tabela de fatos, as tabelas de dimensões e as medidas (MACHADO, 2008).

Em um modelo dimensional a tabela de fatos é a tabela principal que possui enorme volume de dados, constituídos por valores de medidas e códigos de chaves estrangeiras das tabelas de dimensões (REIS, 2009).

As medidas correspondem às variáveis numéricas de um fato, são relativas em relação a cada dimensão e através dela é possível calcular os resultados dos indicadores de negócio. Em um contexto onde o fato é determinado pelas vendas, por exemplo, as medidas que podem existir são: a quantidade de vendas realizadas, o valor total faturado, custo, lucro, entre outras (MACHADO, 2008).

A tabela fato é constituída de vários índices, podendo até ser indexada em todas as colunas, ao indexar todos os registros a tabela fato passa a ter maior viabilidade de acesso. Os dados constantes nas tabelas de fato não são atualizados, assim todos os registros são armazenados de forma incremental (INMON, 1997).

Para melhor entendimento, na Figura 6 a tabela fato foi definida a partir dos dados de vendas.

Figura 6: Tabelas de dimensões e Tabela de fato.

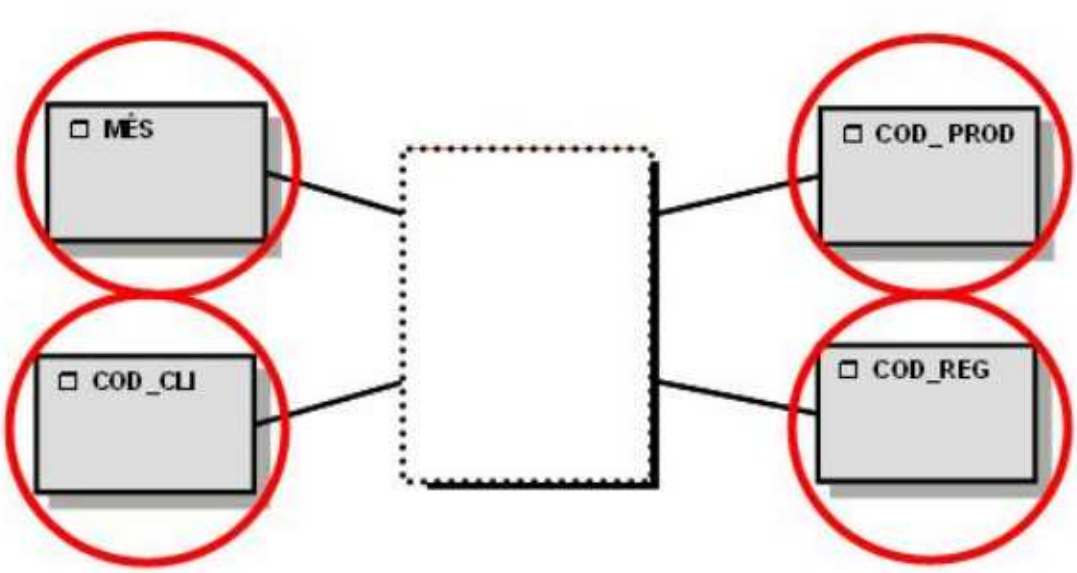
Fonte: Rocha (2007).

Representando características do fato, as tabelas de dimensões contêm dados que permitem realizar diversas análises de um fato, permitindo aos usuários a

apuração de indicadores de maneira detalhada, onde é possível obter o resultado alcançado para cada variável existente no fato. O relacionamento com a tabela de fato é determinado através da chave primária, mantendo integridade referencial (REIS, 2009).

Na figura 7 é possível identificar dimensões de tempo, clientes, produtos e localização, através dos campos MÊS, COD_CLI, COD_PROD e COD_REG respectivamente.

Figura 7: Tabelas de dimensões



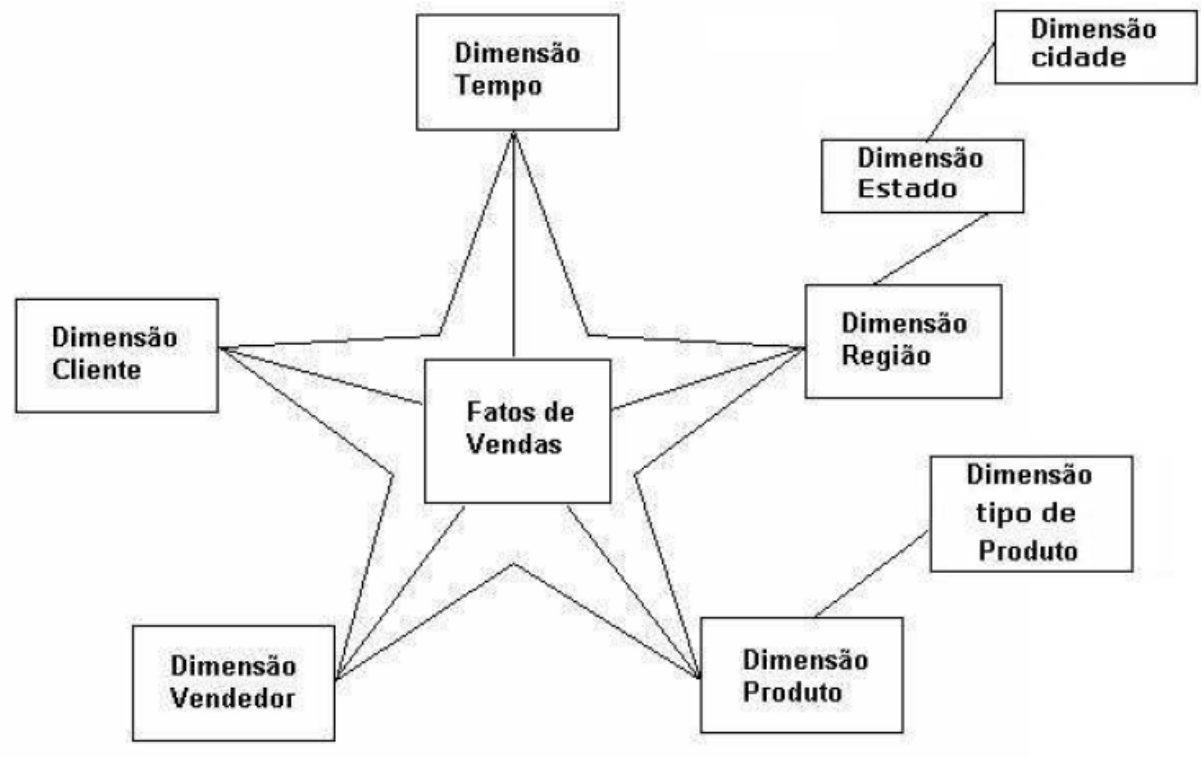
Fonte: Rocha (2007).

Existem dois modelos dimensionais recomendados, o Snowflake (Floco de Neve) e o Star Squema (Estrela).

Quando se tem como objetivo eliminar ao máximo a redundância de dados, o modelo Snowflake é indicado, pois ele é disposto no formato de uma estrela possuindo ramificações nas tabelas de dimensão. As ramificações das dimensões são dispostas de acordo com sua hierarquia. A grande vantagem é a economia de espaço em armazenamento, entretanto, o desempenho nas consulta pode ser prejudicado de acordo com o volume de dados existentes (MACHADO, 2008).

Na Figura 8, a tabela de fato é determinada pelos registros de vendas, contendo dimensões de tempo, clientes, vendedores, produtos e localização. A dimensão de localização é dividida em região, estado e cidade; e a dimensão de produto é subdividida em tipo de produto.

Figura 8: Modelo Snowflake.

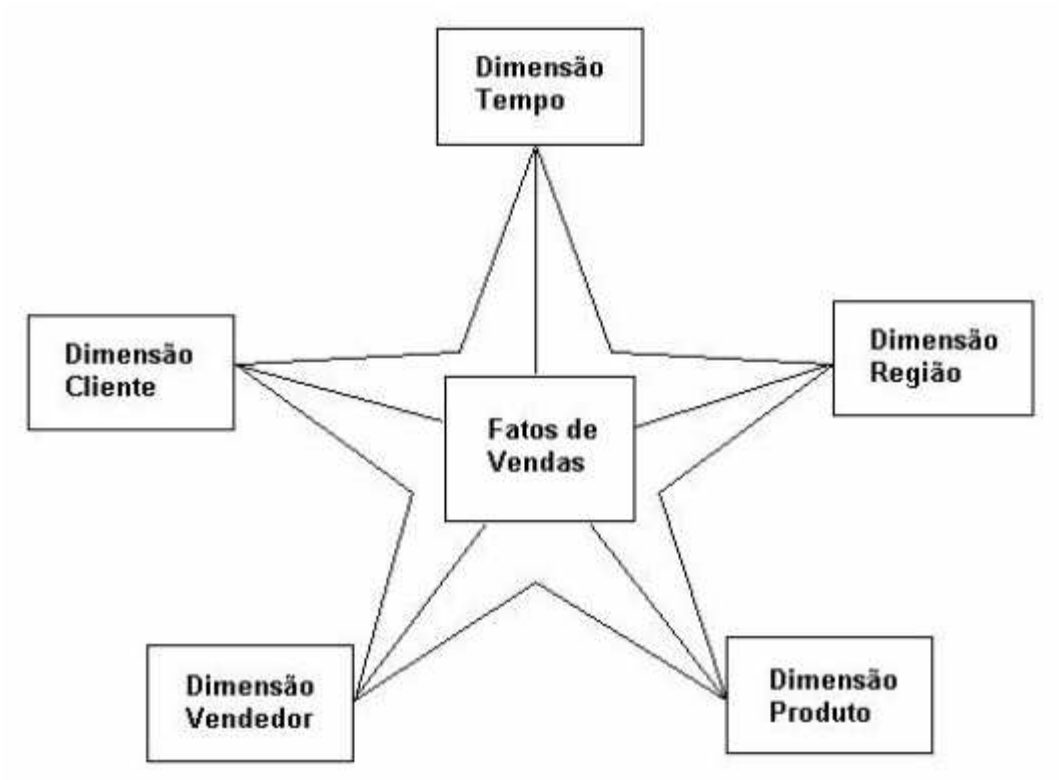


Fonte: Machado (2008).

Se o desempenho em relação ao tempo de consulta é o ponto principal, o modelo Star Schema (estrela) é recomendado pelo fato de possuir diversas tabelas de dimensões ligadas a tabela de fatos. Assim, as dimensões não são divididas em demais tabelas, por este motivo que esta estrutura possui maior número de dados redundantes, exigindo grande espaço para armazenamento (MACHADO, 2008).

Na Figura 9 o modelo Star Schema é representado pela tabela de fatos de vendas e as dimensões de tempo, clientes, vendedores, produtos e localização.

Figura 9: Modelo Star Schema.



Fonte: Machado (2008).

Concluído as etapas que definem a estruturação, os dados devem ser extraídos, padronizados e carregados no Data Warehouse, este processo chama-se ETL e seu conceito foi descrito no item a seguir.

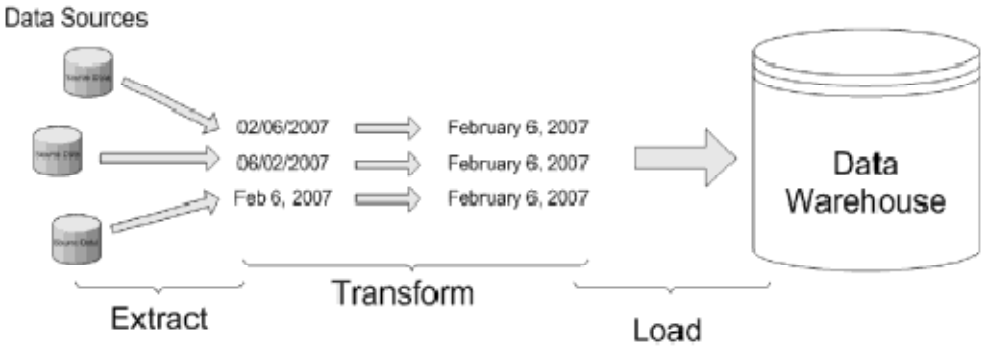
1.3 ETL (Extract Transform Load)

O processo de ETL é fundamental para a criação do alicerce de um sistema de Business Intelligence (BI), ele reúne os dados de diversos sistemas, padroniza-os em formatos consistentes e armazena-os em uma única base de dados, o DW (ECKERSON, 2003).

O processo ETL é responsável por até 80% dos recursos destinados a desenvolvimento em um projeto de construção de um Data Warehouse (INMON,1997).

A Figura 10, representa as 3 etapas do processo ETL.

Figura 10: Processo ETL.



Fonte: Withee (2010).

Antes de iniciar o processo de ETL deve-se definir o nível da granularidade dos dados, essa é uma das principais questões do projeto de um Data Warehouse, ela está relacionada diretamente com a quantidade de dados que serão armazenados e a velocidade das consultas. As organizações podem desejar informações detalhadas, mas deve-se analisar o volume de dados gerados em cada nível de detalhe para definir a melhor opção (INMON, 1997).

Caso o espaço para armazenamento não seja escasso, a melhor opção é guardar os dados em dois níveis de detalhe, assim as organizações podem realizar qualquer tipo de análise, sem comprometer a eficiência das consultas (INMON, 1997).

A extração é a ação de copiar dados de sistemas que armazenam as informações transacionais, estes dados são convertidos para o formato de dados do Data Warehouse, antes do processo de transformação.

O primeiro passo da etapa de extração é definir as fontes de dados que serão utilizadas. Os dados desejados podem advir de vários formatos de armazenamento, entre eles, dados gerenciados através de sistemas de gerenciamento de banco de dados (SGBD) ou até mesmo em um simples arquivo de texto (KIMBALL, 2002).

Devido aos diversos sistemas de armazenamento de dados existentes dentro de uma corporação, desenvolvidos por diferentes fornecedores de tecnologia da informação, é normal que existam dados em formatos incompatíveis. Assim se faz necessário o processo de transformação a que convém padronizar ou limpar dados que sejam inconsistentes (COREY, 2001).

Um exemplo simples é a definição das siglas nas fontes de dados para os sexos masculino e feminino, em um sistema pode-se utilizar a sigla ‘M’ para o sexo masculino e a sigla ‘F’ para feminino, em outro a letra ‘H’ (homem) para masculino e a letra ‘M’ (mulher) para feminino, essa conversão é realizada no processo de transformação, com objetivo de garantir que não ocorram distorções de resultados nas análises realizadas pelos usuários do DW.

Na Tabela 1, são demonstradas as possíveis conversões do atributo ‘sexo’.

Tabela 1: Transformação de atributo

Sistemas Transacionais	Data Warehouse
0 = (masculino)	Masculino
1 = (feminino)	Feminino
H = (homem)	Masculino
M = (mulher)	Feminino
M = (masculino)	Masculino
F = (feminino)	Feminino

Fonte: do autor.

A carga de volumes expressivos de dados é um dos grandes desafios, dependendo da forma como o programador elabora a rotina, esta etapa pode economizar tempo significativo. Um ponto importante para esta atividade é gerenciar corretamente os índices associados às tabelas, os índices melhoram o desempenho das consultas, mas prejudicam o tempo de inserção de novos dados, assim, recomenda-se a remoção dos índices antes de iniciar o processo de carga e a inclusão após o termino do processo (KIMBALL, 2004).

A atualização do Data Warehouse pode ocorrer através da exclusão e inserção completa dos dados ou de forma incremental, atualizando os registros existentes e inserindo novos. Entretanto, as modificações dos registros devem ser evitadas, pois exigem maior capacidade de processamento, portanto, a forma incremental deve ser utilizada apenas para tabelas que não necessitam de alterações dos registros inseridos anteriormente. Assim, a escolha da técnica de carga deve considerar o número de registros a serem atualizados e incrementados, devendo-se realizar simulações para garantir a excelência do processo (KIMBALL, 2004).

1.4 Ferramentas de ETL

Existem diversas ferramentas que auxiliam parcialmente ou totalmente na execução do processo de ETL, ao fazer uso delas é possível afirmar que vários benefícios são agregados ao projeto, com destaque para o aumento significativo da produtividade (PRIMAK, 2009).

Conforme Aguiar (2010), as ferramentas de ETL podem proporcionar os seguintes benefícios:

- Criação de rotinas automatizadas de cargas: define-se quando a carga será iniciada e quais os dados serão carregados.

- Manutenção de cargas: maior facilidade e segurança nas alterações em comparação com a manutenção de códigos.

- Melhor desempenho ao realizar todas as etapas do processo: proporcionada pela existência de métodos que gerenciam os grandes volumes de dados.

- Execução simultânea: extrações, transformações e cargas podem ser executadas em paralelo.

- Acessibilidade: as ferramentas podem acessar arquivos ou realizar cargas, em diversos servidores.

- Conexões: é possível realizar a leitura de qualquer tipo de arquivo e convertê-lo em um novo formato de forma simples e eficaz.

- Etapas por módulos: garante a segurança do processo e diminui a possibilidade de erros.

Existem diversas ferramentas especializadas para realizar as atividades de ETL, dentre as ferramentas comercializada destaca-se o Oracle Warehouse Builder (OWB) e o Microsoft SQL Server Integration Services (SSIS), e entre as ferramentas Open Source o Talend Data Integration (TDI) e o Pentaho Data Integration (PDI). Estas ferramentas possuem características diferenciadas e foram utilizadas em casos de sucesso que serão apresentados a seguir.

O Oracle Warehouse Builder (OWB) é uma ferramenta da Oracle que permite acesso aos dados de sistemas distintos, suporta a movimentação de dados em massa, transformação, carga e gerenciamento da qualidade dos dados. Possui ótimo desempenho para execução das tarefas ETL, executa o processo em sequência diferenciada aos seus concorrentes, inicia-se com a extração, em seguida realiza a carga e por último a transformação dos dados.

Outras funcionalidades do OWB:

- mapeamento de fluxo, além de servir como documentação, é possível aproveitar todo o fluxo desenvolvido ou parte em outros projetos.

- melhor desempenho no processamento devido a execução em paralelo das etapas de integração.

- possui suporte a diversas ferramentas da Oracle para gestão da informação.

Como exemplo de caso de sucesso, Santos (2006) utilizou o OWB para definir, criar e executar rotinas de carga dos dados, implementando um Data Warehouse para auxiliar as tomadas de decisões da Secretaria de Estado de Saúde de São Paulo. Durante o projeto, além das atividades de ETL, foram necessárias a criação de um modelo de dados, a elaboração do projeto físico de banco de dados e a realização do mapeamento das fontes de dados. Após a conclusão do projeto, averiguou-se que todas as etapas do processo de ETL foram executadas com sucesso e que a ferramenta Oracle Warehouse Builder atendeu as expectativas.

O OWB não é comercializado atualmente, uma nova versão foi desenvolvida e o nome atual da ferramenta vendida pela Oracle é Oracle Data Integration que possui novas funcionalidades que suportam todo o processo de ETL, além de possuir melhor desempenho (ORACLE, 2010).

Concorrente direto do OWB, o Microsoft SQL Server Integration Services (SSIS) é um componente visual de fácil utilização dentro da plataforma da Microsoft SQL Server. Esta ferramenta permite a criação de pacotes para importação de dados de diferentes origens, transformação dos dados aplicados às regras de negócio e por fim a carga dos dados em um ou mais destinos.

Disponível a partir da versão Standard do SQL Server, contém recursos mais avançados nas versões Enterprise e Datacenter. Permite a importação de dados de diversas origens, como exemplo, arquivos de texto, Oracle, DB2, MySQL, entre outros.

Outras funcionalidades do SSIS:

-Acessar e distribuir arquivos através de operação de FTP (Protocolo de Transferência de Arquivos);

- Executar comandos SQL, permitindo a adequação dos dados, conforme as regras de negócio;

- Enviar mensagens de e-mail, nas diversas etapas do processo, contendo detalhes das operações executadas;

As versões Standard, Enterprise e Datacenter são comercializados pela Microsoft e estão disponíveis gratuitamente apenas para uso de avaliação por um período de 180 dias (MICROSOFT, 2014).

O SSIS foi utilizado por Pelito (2012) em um estudo de caso onde existia a necessidade de construir um Data Warehouse para análise do processo de coleta de sangue e de medula óssea. O projeto teve como objetivo agrupar as informações existentes nos sistemas transacionais que suportam o processo, reunindo os dados em uma base de dados especializada para análise de informações estatísticas na qual o usuário deste recurso conseguisse realizar o mapeamento das fases deste processo, a fim de conseguir planejar campanhas para diminuição da oscilação do estoque.

No desenvolvimento do projeto de Pelito (2012) o SSIS foi utilizado especificamente para as atividades de ETL, que reunindo as informações de 3 arquivos de formatos diferentes em uma base de dados devidamente modelada, possibilitou diversas análises. Em destaque:

- O monitoramento do consumo médio de sangue e do volume de doações;

- Construção de relatórios estatísticos para conhecimento dos diversos grupos de doadores;

- Mapear os hospitais identificando a qualidade de captação e aproveitamento das bolsas coletadas;

- Identificar o motivo de doadores inaptos e as chances de recuperação dos mesmos;

- Emissão de relatório para acompanhamento dos doadores inaptos em períodos de campanha de vacinação do Governo Federal;

- Acompanhamento dos resultados obtidos em campanhas de incentivo;

- Monitoramento da volumetria de bolsas em estoque por tipo sanguíneo;

- Comparação do número de doadores de medula óssea cadastrados com o total de doações.

Em alternativa às ferramentas que não possuem licenças gratuitas, existem as ferramentas Open Source, o Talend Data Integration é uma delas. Desenvolvida para integração de dados, através dela é possível efetuar todas as etapas de um processo ETL, além disso, é possível realizar o envio de e-mail mediante a conclusão de etapas do fluxo criado.

O TDI conta também com o apoio de uma comunidade de usuários e até mesmo suporte pago. Outras funcionalidades desta ferramenta:

- Desenvolvimento com auxilio de módulos assistentes, gerando um diagrama funcional.

- Possui mais de 450 pacotes de conexão, dentre eles os principais formatos de armazenamento ou sistemas de gerenciamento de banco de dados.

- Suporta conexão mainframes, serviços web e aplicativos em nuvem.

O Talend Data Integration está disponível em três versões, sendo a versão Talend Open Studio for Data Integration gratuita (Open Source) e as versões Talend Enterprise Data Integration e Talend Platform for Data Management comercializadas possuindo mais funcionalidades que apóiam o processo de ETL (TALEND, 2014).

De acordo com o estudo de caso desenvolvido por Tavares (2013) o TDI é utilizado pela empresa Unitel T+ Telecomunicações de Cabo Verde nos processos de ETL, neste projeto foram descritos os principais aspectos da ferramenta, o processo de implementação, as soluções de problemas e os resultados obtidos. Em consequência verificou-se que a ferramenta de ETL Talend é utilizada por fornecer uma interface intuitiva, por ser Open Source, e possuir uma grande variedade de componentes que facilitam o desenvolvimento do processo, além de possuir um grande número de conectores para diferentes Bases de Dados.

Outra opção Open Source é o Pentaho Data Integration, conhecido também como Kettle, nesta ferramenta o processo de extração, transformação e carga de dados é realizado a partir da definição de Metadados em uma plataforma de desenvolvimento visual e intuitiva O PDI conecta-se a qualquer tipo de dado, tem processamento em paralelo, armazenamento em memória de rápido acesso, possibilita a automatização dos processos ETL e efetua o envio de email em qualquer etapa do processo.

Um dos componentes da suíte Pentaho BI, o Pentaho Data Integration está disponível em duas versões. A versão EE (Enterprise Edition) possui parte do código fechado e é comercializada pela Pentaho Corporation que além de oferecer mais funcionalidades, oferece também todo o suporte necessário para utilização da ferramenta. Em compensação, a versão CE (Community Edition) possui todas as funcionalidades necessárias para construção e manutenção de um sistema de BI, seu código é aberto e a ferramenta é mantida por voluntários.

Conforme o estudo de caso desenvolvido por Melo (2010) o Pentaho Data Integration foi utilizado no processo de ETL para construção de um armazém de dados com objetivo de centralizar os dados de pessoas vinculadas à Universidade

Federal da Bahia, criar relatórios estatísticos para auxiliar a tomada de decisões e garantir a veracidade das informações. Neste projeto foram criadas as devidas transformações para alimentar o DW de forma automatizada com dados atualizados em D-1 (dados existentes no bando de dados de produção no dia anterior). Ao término, verificou-se que a utilização do PDI é uma alternativa viável para construção de um Data Warehouse, sendo uma boa opção para empresas com orçamentos reduzidos que desejam criar uma estrutura de BI.

2 Estudo de Caso: Construção de DW com os dados do PAC

Este estudo de caso tem como objetivo demonstrar na prática alguns métodos para construção de um DW, aplicando os conceitos pesquisados durante o desenvolvimento deste trabalho e também serão apresentados os resultados obtidos na execução de consultas, comparando assim o desempenho da estrutura de um Data Warehouse e de uma estrutura convencional.

Em virtude da pesquisa realizada, o sistema Pentaho Data Integration foi o sistema de suporte para as atividades de ETL. Este sistema foi escolhido por ser Open Source, por possuir as funcionalidades de extração, transformação e carga de dados, atendendo à necessidade deste estudo de caso. O PDI foi escolhido também por ser uma ferramenta que pode ser executada sem instalação, ou seja, basta descompactar os arquivos para utilizá-la, além do fato de possuir uma interface gráfica intuitiva, orientada a Metadados, permitindo que os métodos estudados sejam testados sem grandes dificuldades.

A base de dados utilizada é a do PAC (Programa de Aceleração do Crescimento) que é disponibilizada e atualizada pelo Governo Federal e de livre acesso por qualquer pessoa ou organização. Esta base foi escolhida por possuir características que permitem a criação de uma tabela de fatos e as respectivas tabelas de dimensões. Esta base está contida em arquivos com formatos de dados diferentes, similar ao contexto existente em diversas corporações.

2.1 Fonte de dados do PAC

Os arquivos com os dados do PAC foram obtidos no repositório do governo, como demonstrado na Figura 11.

Figura 11: Repositório de Dados do PAC.

Fonte: Repositório de Dados do PAC.

Os principais dados sobre os empreendimentos das obras do PAC serão extraídos dos arquivos ‘PAC_YYYY_MM.xml’, os dados destes serão utilizados para popular a tabela de fato e tabelas de dimensões. A partir das informações existentes no arquivo ‘DICIONARIO_DE_DADOS.pdf’ foram gerados outros dois arquivos, ‘ESTAGIO.txt’ e ‘DIGS.csv’, que possuem respectivamente dados das descrições de estágio e tipo de empreendimento, que serão inseridos nas tabelas de dimensões.

Os arquivos ‘ESTAGIO.txt’ e ‘DIGS.csv‘ foram gerados em formatos diferentes com o objetivo de representar as diferentes fontes de dados existentes em uma corporação.

Nas tabelas abaixo são apresentados os nomes, tipos e descrições dos dados que compõem cada arquivo. Na Tabela 2 o arquivo ‘PAC_YYYY_MM.xml’, na Tabela 3 o arquivo ‘DIGS.csv‘ e na Tabela 4 o arquivo ‘ESTAGIO.txt’.

Tabela 2: Dados que compõem o arquivo 'PAC_YYYY_MM.xml'.

Campo	Tipo	Descrição
idn_empreendimento	Inteiro	Identificador único de cada empreendimento.
idn_digs	Inteiro	Identificador do tipo e subeixo de cada empreendimento. Vide seção "Conversão Digs".
dsc_titulo	Texto	Nome do empreendimento.
val_2011_2014	Decimal	Valor em Reais do investimento previsto para o empreendimento entre 2011 e 2014.
val_pos_2014	Decimal	Valor em Reais do investimento previsto para o empreendimento depois de 2014.
investimento_total	Decimal	Utilizado em substituição aos dois campos anteriores nos empreendimentos das áreas Social e Urbana, cujos investimentos não são separados em "PAC 1" e "PAC 2".
sig_uf	Texto	Estado(s) onde o empreendimento está localizado.
txt_municipios	Texto	Município(s) onde o empreendimento está localizado.
txt_executores	Texto	Órgão ou entidade pública ou privada responsável pela execução do empreendimento.
dsc_orgao	Texto	Órgão superior responsável pelo acompanhamento e monitoramento do empreendimento.
idn_estagio	Inteiro	Identificador do estágio do empreendimento. Vide seção "Conversão Estágio".
dat_ciclo	Data	Data do ciclo (processo básico de monitoramento). Data limite de atualizações das informações do empreendimento.
dat_selecao	Data	Data em que o empreendimento foi selecionado e incluído na carteira de projetos do PAC.
dat_conclusao_revisada	Data	Data de conclusão do empreendimento atualizada e revisada.
val_lat	Texto	Latitude do centro do município do empreendimento no formato DMS (degree, minute, second). Para empreendimentos localizados em mais de um município se utilizou um valor médio.
val_long	Texto	Longitude do centro do município do empreendimento no formato DMS (degree, minute, second). Para empreendimentos localizados em mais de um município se utilizou um valor médio.
Emblemática	Texto	Os principais empreendimentos da carteira de projeto do PAC, do ponto de vista da materialidade, relevância ou impacto, recebem o valor "EMBLEMATICA".
Observação	Texto	Observações diversas sobre os empreendimentos.

Fonte: próprio.

Tabela 3: Dados que compõem o arquivo ‘DIGS.csv’.

Campo	Tipo	Descrição
idn_digs	Inteiro	Identificador único de cada tipo de empreendimento
Subeixo	Texto	Grupo do tipo de empreendimento
Tipo	Texto	Sub-Grupo do tipo de empreendimento

Fonte: próprio.

Tabela 4: Dados que compõem o arquivo ‘Estagio.txt’.

Campo	Tipo	Descrição
idn_estagio	Inteiro	Identificador único de cada estágio do empreendimento
Estagio	Texto	Estágio do empreendimento
Descrição	Texto	Descrição do estágio do empreendimento

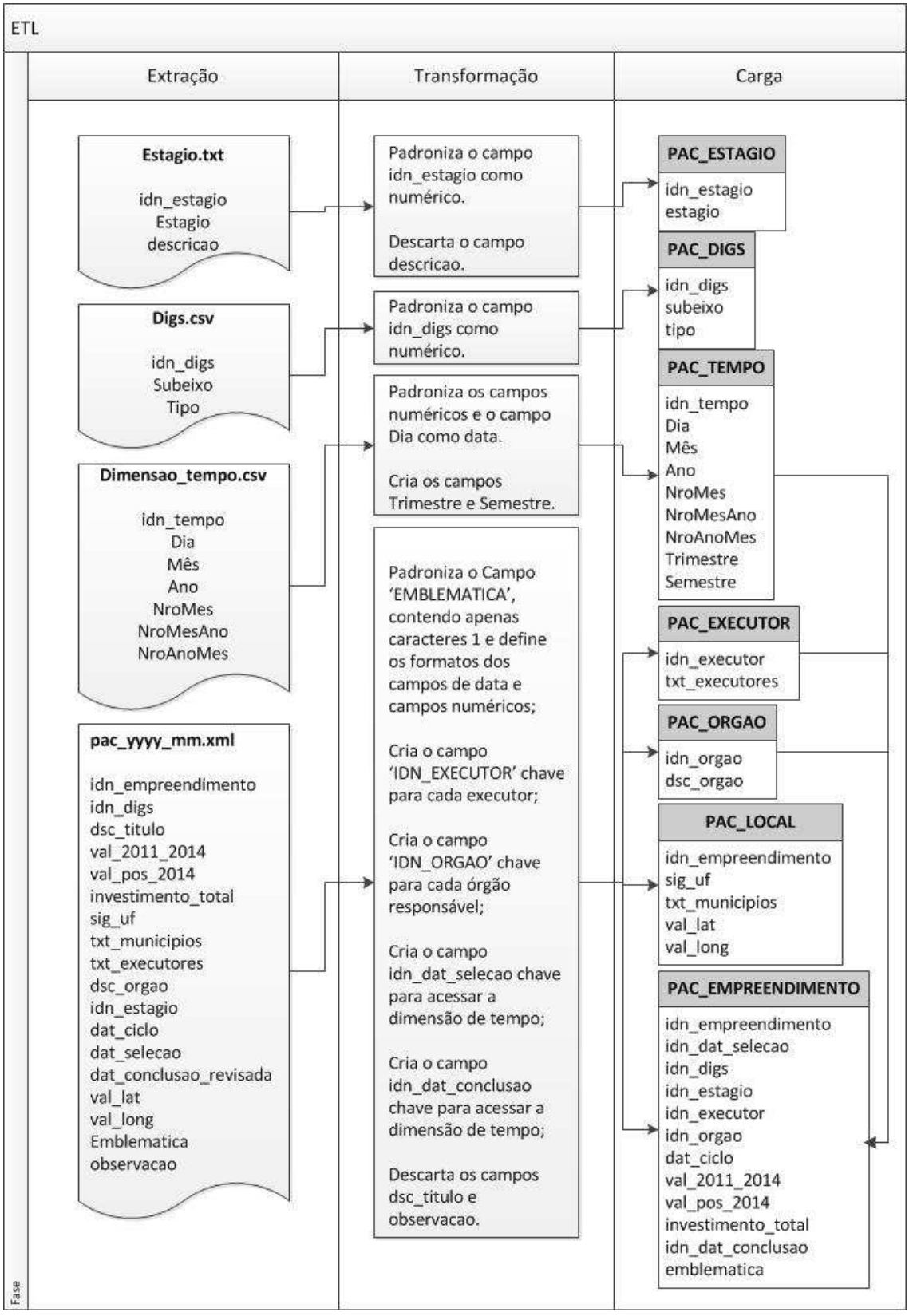
Fonte: próprio.

2.2 Desenvolvimento do Fluxo ETL e da Modelagem Dimensional

Conforme revisão bibliográfica, para fazer a análise, o modelo utilizado neste projeto foi o Star Schema pelo fato de possuir melhor desempenho de processamento de dados, permitindo assim uma entrega ágil de informações. O fato deste método exigir maior espaço de armazenamento, foi considerado menos relevante, pois atualmente o custo de dispositivos para esta finalidade está reduzindo a cada dia.

Na próxima página o fluxograma representa o processo de ETL, onde é possível visualizar todo o fluxo dos dados realizado para compor o Data Warehouse. O primeiro bloco contém os arquivos de onde serão extraídas as informações, no segundo as devidas transformações necessárias para o modelo Star Schema e por último o bloco contendo as tabelas de fato e dimensões.

Figura 12: Fluxograma do processo ETL.

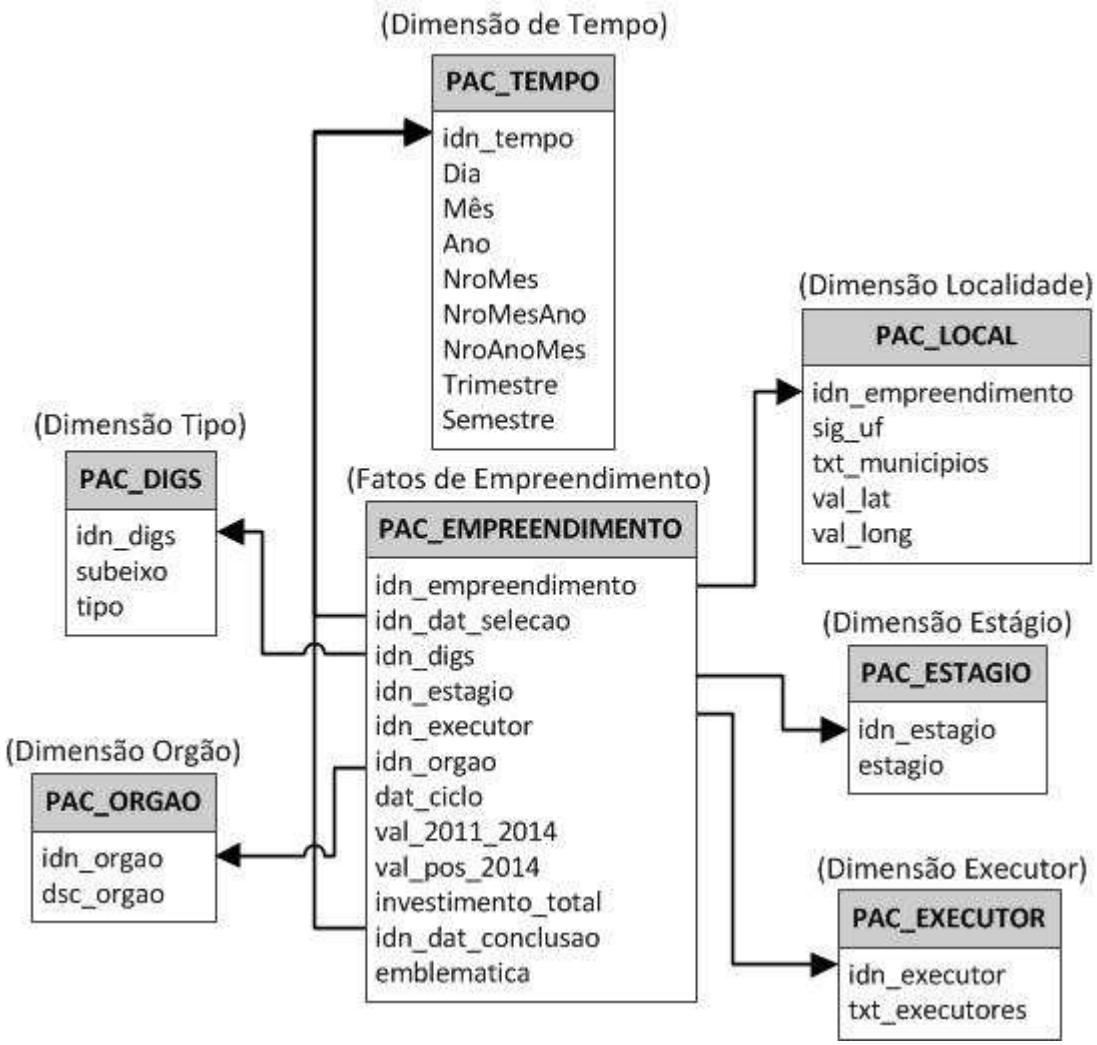


Fonte: próprio.

Após a conclusão do processo de ETL, os dados ficam dispostos nas tabelas do Data Warehouse, onde a tabela ‘PAC_EMPREENDIMENTO’ representa o fato do empreendimento; a tabela ‘PAC_TEMPO’, a dimensão de tempo; a tabela ‘PAC_LOCAL’ a dimensão de localidade, a tabela ‘PAC_EXECUTOR’, a dimensão de executor; a tabela ‘PAC_ORGAO’, a dimensão de órgão; a tabela ‘PAC_ESTAGIO’, a dimensão de estágio; e por fim a tabela ‘PAC_DIGS’ representa a dimensão de tipo.

A Figura 13 apresenta o relacionamento entre as tabelas de fato e dimensões do Data Warehouse.

Figura 13: Entidades e relacionamento do DW.



Fonte: do autor.

O sistema gerenciador de banco de dados utilizado para o desenvolvimento deste estudo foi o MySQL. Este sistema foi escolhido por ser Open Source, e por ser utilizado por várias empresas e pelo fato de ser compatível com as funcionalidades do Pentaho Data Integration.

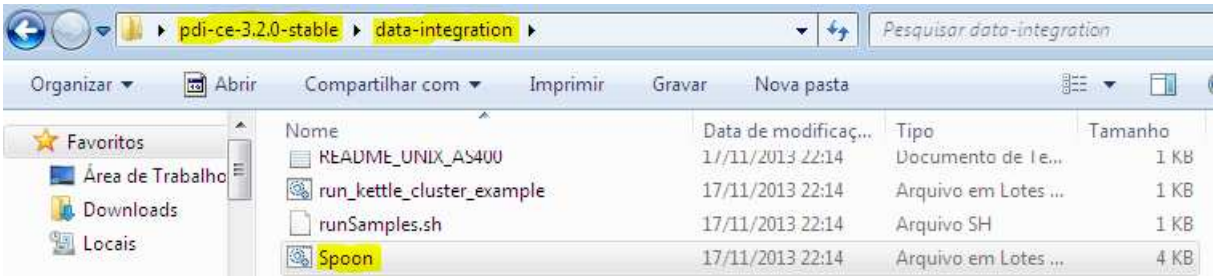
O código SQL do Apêndice A foi desenvolvido para criar as devidas tabelas no Data Warehouse.

Após a criação do DW os dados serão inseridos no banco de dados novamente, porém na estrutura original dos arquivos do PAC, com o intuito de simular a base de dados de um sistema em produção, para que seja possível comparar a efetividade da modelagem dimensional desenvolvida. O Apêndice B possui o script desenvolvido para criação das tabelas utilizadas para simular consultas na estrutura original dos arquivos.

2.3 Execução da ETL com o Pentaho Data Integration

Após a criação das tabelas no banco de dados, foi necessário efetuar a instalação do Java Runtime Environment (JRE), obtido no site ‘http://Java.com’ e posteriormente descompactar no diretório do estudo de caso o arquivo do sistema Pentaho Data Integration, versão 3.2.0. Na sequência, o PDI foi utilizado executando-se o arquivo ‘Spoon.bat’ disponível no diretório pdi-ce-3.2.0-stable\data-integration, de acordo com a Figura 14.

Figura 14: Execução Kettle.



Fonte: do autor.

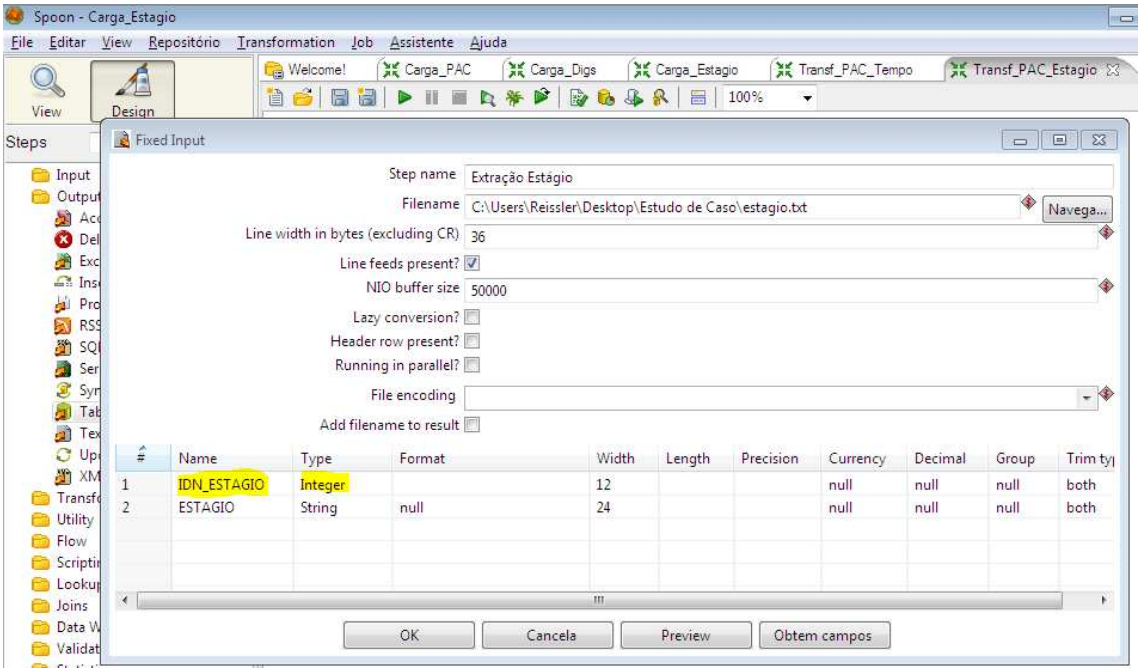
Ao executar o Spoon a tela inicial do PDI é exibida e o menu de desenvolvimento é aberto após clicar no botão ‘No repository’, conforme a Figura 15.

Figura 15: Tela inicial Kettle.

Fonte: do autor.

O processo de ETL foi iniciado com a extração do arquivo ‘Estagio.txt’, na etapa de transformação o campo ‘idn_estagio’ foi padronizado no formato numérico e o campo descrição foi descartado pelo fato de não ser uma informação relevante para análise gerencial, conforme a Figura 16.

Figura 16: Processo ETL - Extração e Padronização do campo ‘idn_estagio’.



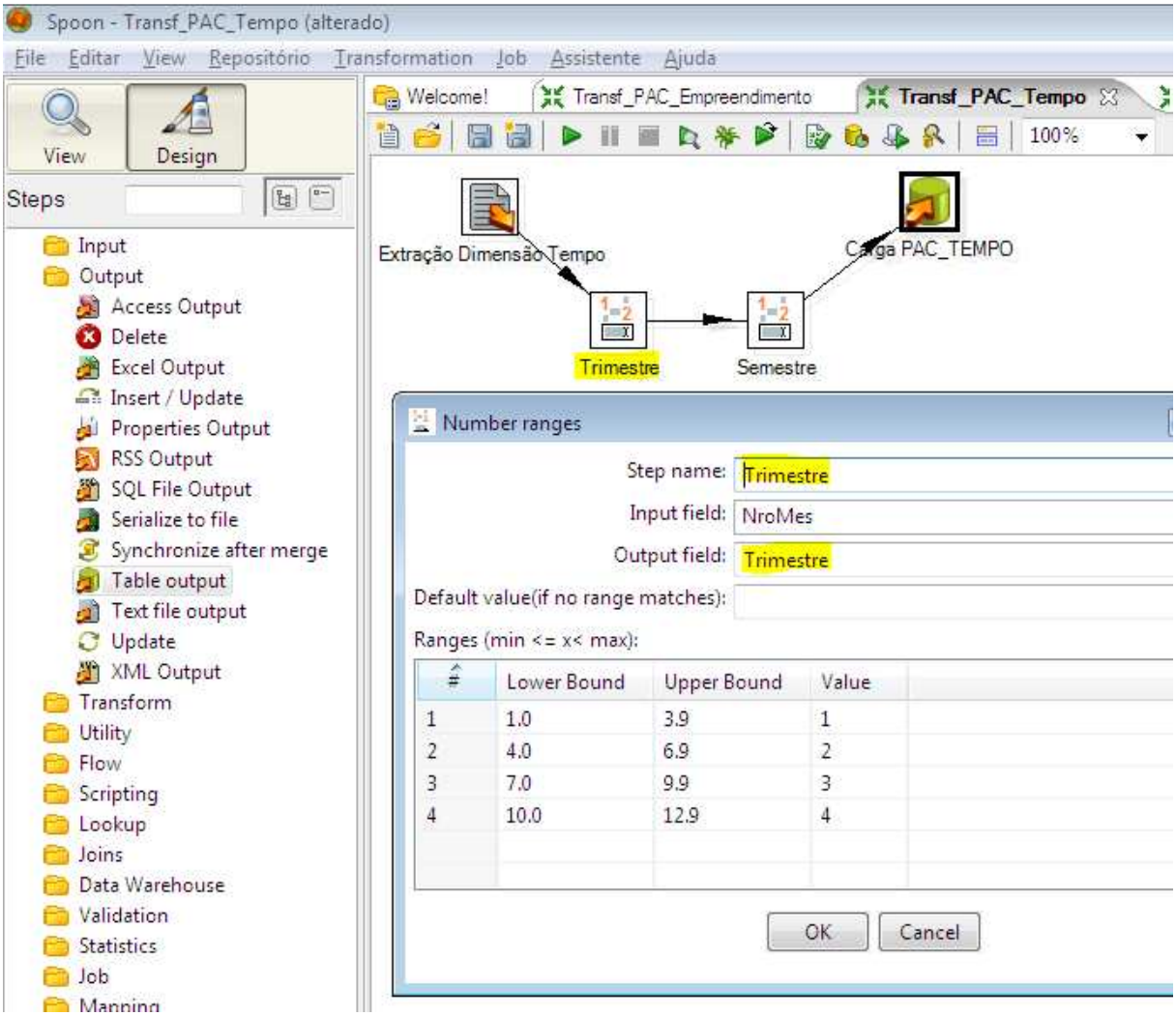
Fonte: do autor.

Os dados foram carregados na tabela PAC_ESTAGIO do Data Warehouse.

Em seguida foi realizada a extração do arquivo ‘Digs.csv’, na etapa de transformação o campo ‘idn_digs’ foi padronizado como numérico e os dados foram carregados na tabela PAC_DIGS do DW.

Ao prosseguir com o estudo de caso os dados do arquivo ‘Dimensao_tempo.csv’ foram extraídos. Na transformação dos dados foram realizadas as padronizações dos campos numéricos (‘idn_tempo’, ‘ano’, ‘nromes’, ‘nromesano’ e ‘nroanomes’) e do campo ‘dia’ no formato de data, além disso, foram criados os campos ‘trimestre’ e ‘semestre’ com base nas informações contidas no campo ‘nromes’, conforme a Figura 17.

Figura 17: Processo ETL - Dimensão de tempo, criação do campo ‘trimestre’.

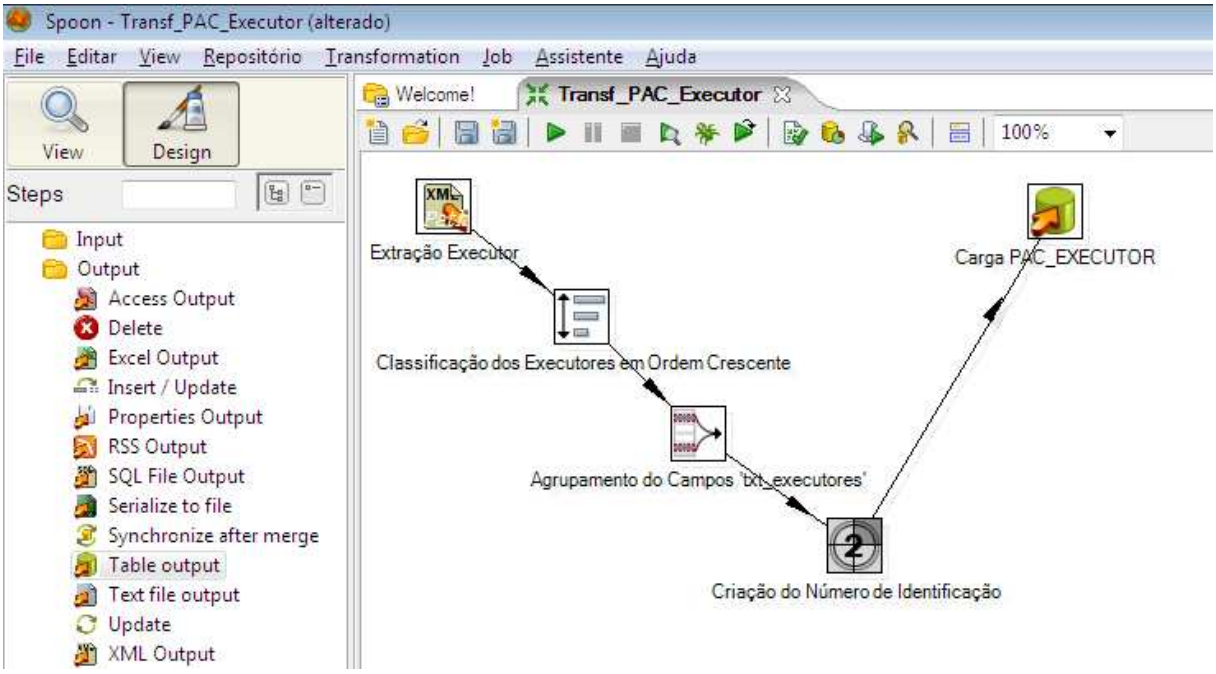


Fonte: do autor.

Após a extração e transformação, a carga foi realizada na tabela PAC_TEMPO do Data Warehouse.

Na sequência foi realizado o processo ETL para carregar os dados nas tabelas PAC_EXECUTOR e PAC_ORGAO, na transformação os dados foram consolidados para identificar os registros distintos, consecutivamente foram criados os campos ‘idn_executor’ e ‘idn_orgao’ chave para localização de cada executor e órgão. Na Figura 18 é apresentado o processo ETL desenvolvido na ferramenta do Pentaho Data Integration para execução da etapa de ETL dos executores.

Figura 18: Processo ETL - Executores.



Fonte: do autor.

A fase de carga da tabela PAC_LOCAL foi desenvolvida com a extração do arquivo ‘PAC.xml’, na transformação apenas o campo ‘idn_empreendimento’ foi padronizado como numérico e não foi criado um novo campo de identificação para cada local, pois foi considerado inviável, devido este campo possuir agregações de vários municípios (conforme a Figura 19), sendo que a criação de um novo número de identificação resultaria em quantidade aproximada de identificadores em relação ao identificador idn_empreendimento.

Figura 19: Dados existentes no campo ‘txt_municipios’.

txt_municipios
GUARULHOS
PORTO VELHO/RO, CANDEIAS DO JAMARI/RO, ALTO PARAÍSO/RO, ARIQUEMES/RO, CACAULÂNDIA/RO, JARU/RO, OURO PRETO DO OESTE/RO, TEIXEIRÓPOLIS/RO, JI-PARANÁ/R(
ORIXIMINÁ/PA, AMAPÁ/AP
ITAPIRANGA/SC, MONDAÍ/SC, SÃO JOÃO DO OESTE/SC, CAIÇARA/RS, PINHEIRINHO DO VALE/RS, VICENTE DUTRA/RS, VISTA ALEGRE/RS

Fonte: do autor.

Por fim, foi realizado o processo de carga dos dados na tabela PAC_EMPREENDIMENTO do DW, nesta fase os dados foram extraídos do arquivo ‘PAC.xml’, sendo que na etapa de transformação foram realizadas as padronizações, conforme a Tabela 5.

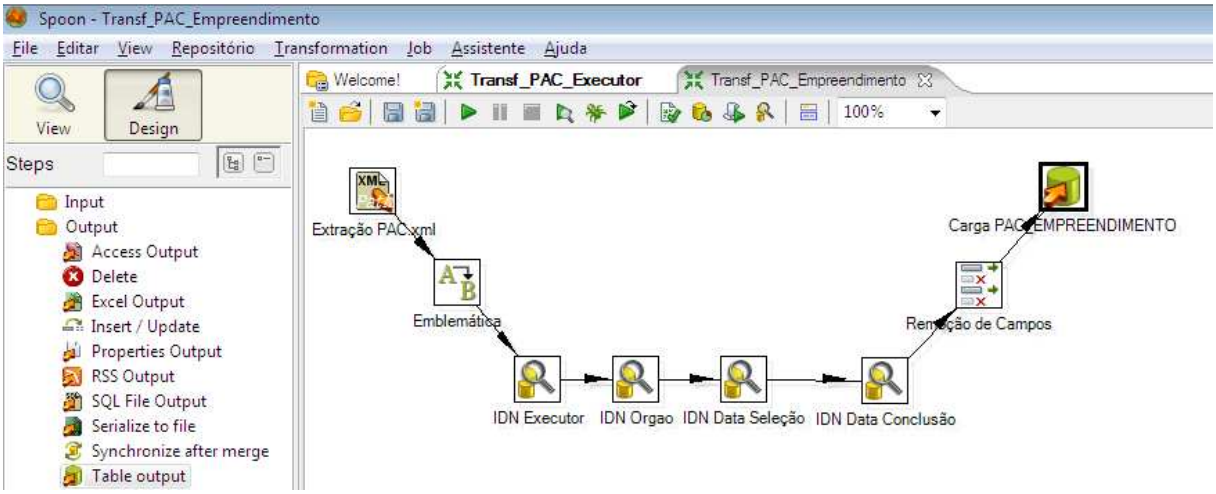
Tabela 5: Transformações tabela de Fato

Campo	Transformação
idn_empreendimento	Numérico
idn_digs	Numérico
idn_estagio	Numérico
val_2011_2014	Numérico
val_pos_2014	Numérico
investimento_total	Numérico
dat_ciclo	Data
Emblemática	Numérico (1 ou 0)
idn_executor	Incrementado a partir da dimensão de executores
idn_orgao	Incrementado a partir da dimensão de Orgãos
idn_dat_selecao	Incrementado a partir da dimensão de Tempo
idn_dat_conclusao	Incrementado a partir da dimensão de Tempo
dsc_titulo	Removido
Observação	Removido

Fonte: do Autor.

A transformação do campo ‘emblematica’ foi realizada com o objetivo de reduzir o tamanho do campo e consecutivamente melhorar a velocidade de acesso a está informação. A Figura 20 apresenta o processo completo de ETL desta fase.

Figura 20: Processo ETL - tabela de fato (PAC_EMPREENDIMENTO).



Fonte: do autor.

2.4 Testes de Efetividade do DW

Após carga de todos os arquivos do PAC, inclusive os arquivos com dados retroativos, foram executadas três consultas, com objetivo de simular possíveis consultas provenientes de análises de negócio. As execuções foram realizadas desabilitando o maior número de serviços possíveis do sistema operacional, além disso, foram feitas três execuções para cada consulta, com o propósito de diminuir a possibilidade de interferência nos resultados.

Tabela 6: 1ª Consulta Estrutura DW e Estrutura Original

Data Warehouse	Sistema Original
Código SQL:	
select conclusao_mes, count(idn_empreendimento) quantidade, sum(investimento_total) investimento_total from datawarehouse.pac_empreendimento, datawarehouse.estagio, datawarehouse.view_pac_tempo_conclusao where pac_empreendimento.idn_estagio = estagio.idn_estagio and idn_dat_conclusao = conclusao_idn_tempo and conclusao_ano = '2013' and estagio = 'Concluído' and dat_ciclo = '2013-12-31' group by conclusao_mes order by 2;	select case when month(dat_conclusao_revisada) = '1' then 'janeiro' when month(dat_conclusao_revisada) = '2' then 'fevereiro' when month(dat_conclusao_revisada) = '3' then 'março' when month(dat_conclusao_revisada) = '4' then 'abril' when month(dat_conclusao_revisada) = '5' then 'maio' when month(dat_conclusao_revisada) = '6' then 'junho' when month(dat_conclusao_revisada) = '7' then 'julho' when month(dat_conclusao_revisada) = '8' then 'agosto' when month(dat_conclusao_revisada) = '9' then 'setembro' when month(dat_conclusao_revisada) = '10' then 'outubro' when month(dat_conclusao_revisada) = '11' then 'novembro' when month(dat_conclusao_revisada) = '12' then 'dezembro' end mes, count(idn_empreendimento) quantidade, sum(investimento_total) investimento_total from datawarehouse.pac, datawarehouse.estagio where pac.idn_estagio = estagio.idn_estagio and year(dat_conclusao_revisada) = '2013' and estagio = 'Concluído' and dat_ciclo = '2013-12-31' group by month(dat_conclusao_revisada) order by 2;
Tempo de Execução:	
0,562 sec	0,562 sec
0,563 sec	0,546 sec
0,562 sec	0,562 sec

Fonte: do autor.

Figura 21: Dados de retorno da 1ª Consulta.

Mes	Quantidade	investimento_total
Fevereiro	143	9856321706
Novembro	153	11975774204
Outubro	157	153390451942
Janeiro	162	5881037111
Março	170	31945079699
Setembro	172	4631159947
Maio	174	35377797622
Julho	177	31362685094
Junho	189	10586909638
Agosto	193	20959346769
Abril	251	23081136416
Dezembro	282	14893353921

Fonte: do autor.

A primeira consulta foi desenvolvida para retornar a volumetria dos empreendimentos concluídos no ano de 2013 por Mês de conclusão, nesta consulta fica evidente a diferença entre a quantidade de código da consulta realizada nas tabelas do Data Warehouse e da estrutura original. Em relação ao tempo de execução não ouve uma diferença significativa considerando as três execuções efetuadas.

Tabela 7: 2ª Consulta Estrutura DW e Estrutura Original

Data Warehouse	Sistema Original
Código SQL:	
select subeixo, count(*) Quantidade, sum(val_pos_2014) investimento_pos_2014 from datawarehouse.pac_empreendimento, datawarehouse.digs, datawarehouse.estagio where pac_empreendimento.idn_digs = digs.idn_digs and pac_empreendimento.idn_estagio = estagio.idn_estagio and val_pos_2014 is not null and estagio <> 'Concluído' and dat_ciclo = '2013-12-31' group by subeixo order by 3 desc;	select subeixo, count(*) Quantidade, sum(val_pos_2014) investimento_pos_2014 from datawarehouse.pac, datawarehouse.digs, datawarehouse.estagio where pac.idn_digs = digs.idn_digs and pac.idn_estagio = estagio.idn_estagio and val_pos_2014 is not null and estagio <> 'Concluído' and dat_ciclo = '2013-12-31' group by subeixo order by 3 desc;
Tempo de Execução:	
0,515 sec	0,547 sec
0,516 sec	0,563 sec
0,516 sec	0,562 sec

Fonte: do autor.

Figura 22: Dados de retorno da 2ª Consulta.

subeixo	Quantidade	investimento_pos_2014
Energia	136	66250121500000
Transportes	64	1129471210364
Cidade Melhor	20	840535500000
Água e Luz para todos	41	168724884955

Fonte: do autor.

Na segunda consulta foi observado um pequeno ganho de desempenho na estrutura do Data Warehouse em relação a estrutura original. Ao executar a consulta obteve-se a distribuição dos empreendimentos que não foram concluídos e que possuem contratos com investimento para depois do ano de 2014.

Tabela 8: 3ª Consulta Estrutura DW e Estrutura Original

Data Warehouse	Sistema Original
Código SQL:	
select conclusao_trimestre, count(*) quantidade, sum(investimento_total) investimento_total from datawarehouse.pac_empreendimento a, datawarehouse.estagio b, datawarehouse.view_pac_tempo_conclusao c where a.idn_estagio = b.idn_estagio and a.idn_dat_conclusao = c.conclusao_idn_tempo and conclusao_ano = '2013' and estagio = 'Concluído' and emblematica = 'EMBLEMATICA' and dat_ciclo = '2013-12-31' group by conclusao_trimestre order by 1;	select case when month(dat_conclusao_revisada) in ('1','2','3') then '1' when month(dat_conclusao_revisada) in ('4','5','6') then '2' when month(dat_conclusao_revisada) in ('7','8','9') then '3' when month(dat_conclusao_revisada) in ('10','11','12') then '4' end trimestre, count(*) quantidade, sum(investimento_total) investimento_total from datawarehouse.pac, datawarehouse.estagio where pac.idn_estagio = estagio.idn_estagio and year(dat_conclusao_revisada) = '2013' and estagio = 'Concluído' and emblematica = 'EMBLEMATICA' and dat_ciclo = '2013-12-31' group by case when month(dat_conclusao_revisada) in ('1','2','3') then '1' when month(dat_conclusao_revisada) in ('4','5','6') then '2' when month(dat_conclusao_revisada) in ('7','8','9') then '3' when month(dat_conclusao_revisada) in ('10','11','12') then '4' end order by 1;
Tempo de Execução:	
0,516 sec	0,563 sec
0,515 sec	0,562 sec
0,500 sec	0,578 sec

Fonte: do autor.

Figura 23: Dados de retorno da 3ª Consulta.

CONCLUSAO_TRIMESTRE	Quantidade	investimento_total
1	5	1420448138
2	4	25615890535
3	4	9728401218
4	5	118574151500

Fonte: do autor.

Na execução da última consulta os dados retornados referem-se aos empreendimentos emblemáticos concluídos no ano de 2013 separados por trimestre. O tempo da consulta foi menor na estrutura do Data Warehouse do que na estrutura do sistema original e o código SQL utilizado na consulta do DW foi mais compacto do que o código SQL executado nas tabelas do sistema original.

3 Considerações Finais

Diante da grande competitividade empresarial e da assertividade em tomadas de decisões proporcionadas pelas análises de dados, aliado ao crescente volume de dados e as diferentes fontes de armazenamento, se faz necessário a elaboração de uma estrutura que organize e facilite o acesso a informação.

A construção desta estrutura denominada como Data Warehouse, é viabilizada por métodos desenvolvidos por especialistas no assunto. Dentre os métodos destacam-se os Metadados que mapeiam toda a estrutura, a arquitetura de DW e DM que favorece o gerenciamento das permissões de acesso de usuários, a modelagem dimensional que organiza os dados facilitando e agilizando as análises e por fim o processo ETL, responsável pela confiabilidade dos dados.

Além disso, a utilização de uma ferramenta que auxilie no processo ETL é fundamental para simplificar a extração dos dados obtidos em fontes e formatos distintos, auxiliar na normalização e filtragem dos dados, automatizar rotinas e documentar as etapas do processo.

Em conseqüência ao estudo de caso desenvolvido ao longo deste trabalho, foi observado que a utilização de alguns métodos facilitou a elaboração de análises de negócio a partir de consultas SQL, porém o desempenho na execução dessas consultas não obteve ganho significativo pelo fato da base utilizada possuir poucos registros devido a sua atualização mensal, em comparação com a quantidade de registros existentes em bases de dados de grandes organizações que possuem diversos sistemas e transações diárias. Entretanto o pequeno ganho de

desempenho obtido com os métodos aplicados pode ser mais significativo se o volume de dados for maior.

Com a conclusão deste trabalho, estudos futuros podem ser desenvolvidos com base na estrutura criada neste estudo de caso. Pode-se propor o levantamento de técnicas de mineração de dados que permitam a descoberta de correlações que não sejam possíveis através da exploração manual realizadas por usuários, utilizando-se uma ferramenta estatística que auxilie este processo.

O levantamento dos melhores métodos para construção de um Data Warehouse foi significativo para demonstrar aos profissionais de TI e qualquer outro profissional que utilize informações geradas a partir de um DW, a importância do desenvolvimento de uma estrutura qualificada para as análises de inteligência de negócios, em vista que decisões podem acarretar no sucesso ou fracasso de uma organização.

REFERÊNCIAS BIBLIOGRAFICAS

AGUIAR, Gustavo Maia. **Por que utilizar uma ferramenta de ETL?**. Disponível em: < http://gustavomaiaaguiar.wordpress.com/2010/05/10/por-que-utilizar-uma-ferramenta-de-etl/>. Acesso em: 09 nov. 2013.

BARBIERI, Carlos. **BI - Business Intelligence - Modelagem & tecnologia**. Rio de Janeiro: Axcel Books do Brasil Editora, 2001.

BONOMO, Peeter. **Arquitetura de Data Warehouse – Parte 01**. Disponível em: <http://imasters.com.br/artigo/11417/gerencia-de-ti/arquitetura-de-data-warehouse-parte-01/>. Acesso em: 29 set. 2013.

BONOMO, Peeter. **Arquiterura de Data Warehouse – Parte 02**. Disponível em: <http://imasters.com.br/artigo/11721/gerencia-de-ti/arquitetura-de-data-warehouse-parte-02/>. Acesso em: 29 set. 2013.

BONOMO, Peeter. **Construção de Data Warehouse (DW) e Data Mart (DM)**. Disponívelem:<http://imasters.com.br/artigo/11178/gerencia/construcao_de_data_warehouse_dw_e_data_mart_dm/>. Acesso em: 27 set. 2013.

CHOO, C. W. **The management of uncertainty: organizations as decision-making systems**. New York: Oxford University, 1998.

COREY, Michael et al. **Oracle 8i Data Warehouse**. Tradução de João Tortello. Rio de Janeiro: Campus, 2001.

ECKERSON, Wayne, WHITE, Colin. **Evaluating ETL and Data Integration Platforms**. Seattle: The Data Warehousing Institute, 2003.

GONÇALVES, Marcio. **Extração de Dados para Data Warehouse**. Rio de Janeiro: Axel, 2003.

HAMMERGREN, Thomas C.; SIMON, Alan R. **Data Warehousing For Dummies**. 2. Ed. Indianapolis: Wiley Publishing, Inc. 2009.

HOWSON, Cindi. **Successful Business Intelligence: secrets to making BI a killer app**. New York: McGraw-Hill, 2008.

INMON, William H. **Como Construir o Data Warehouse**. Rio de Janeiro: Campus, 1997.

INMON, William. H.; HACKARTHORN, Richard. D. **Como Usar o Data Warehouse**. Rio de Janeiro: IBPI Press, 1997.

KIMBALL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling**. New York: Wiley Computer Publishing, 2002.

KRAEMER, M. E. P. **Capital intelectual: a nova vantagem competitiva**. Disponível em: <http://www.gestiopolis.com/recursos3/docs/ger/capintel.htm>. Acesso em: 23 set. 2013.

MACHADO, Felipe Nery Rodrigues. **Tecnologia e Projeto de Data Warehouse: Uma visão multidimensional**. São Paulo: Érica, 2008.

MELO, Ana Cristina do Espírito Santo; RAMOS, Gustavo; PURIFICAÇÃO, Mauricio Cesar Santos; VIEIRA, Vaninha. **Uma Experiência de Solução de Business Intelligence com Software Livre na UFBA Utilizando a Suíte Pentaho BI**. Disponível em: <http://pt.slideshare.net/mscesar/uma-experincia-de-soluo-de-business-intelligence-com-software-livre-na-ufba-utilizando-a-sute-pentaho-bi> Acesso em: 27 fev. 2014.

MICROSOFT. **Tutorial SSIS**. Disponível em: <http://technet.microsoft.com/pt-br/library/ms169917.aspx> Acesso em 20 fev. 2014.

NARDI, Alexandre Ricardo. **Fundamentos e Modelagem de Banco de Dados Multidimensionais.** 2000. Disponível em**:** <http://msdn.microsoft.com/pt-br/library/cc518031.aspx> Acesso em: 20 out. 2013.

OLIVEIRA, Wilson José de. **Data warehouse**. Florianópolis: Visual Books Ltda, 2002.

ORACLE. **Oracle Warehouse Builder 11gR2 New Features: Data Integration and Data Warehousing**. Disponível em: <http://www.oracle.com/ technetwork/developer-

tools/warehouse/owb-11gr2-new-features-summary-129693.pdf> Acesso em: 08 mar. 2014.

PELITO, Rogério de Torres; PEREIRA, Gleise Celeste Gonzaga; SOUZA, Diana Maria da Silva; OLIVEIRA, André Luiz Alves; ESCOVEDO, Tatiana; MELO, Rubens N. **Uma Aplicação de Data Warehouse para Análise do Processo de Coleta de Sangue e de Medula Óssea**. Disponível em: <http://www.infobrasil.inf.br/userfiles/15-S1-3-97206-Uma%20aplica%C3%A7%C3%A3o%20de%20Data%20Warehouse___.pdf> Acesso em: 20 fev. 2014.

PENTAHO. **Pentaho Data Integration**. Disponível em: <http://www.pentaho.com/product/data-integration> Acesso em: 27 fev. 2014.

PRIMAK, Fábio Vinícius. **Decisões com BI (Business Intelligence)**. Rio de Janeiro: Ciência Moderna, 2008.

REIS, E., TEIXEIRA, F. e ARAÚJO, M. A.. **Implementando uma solução de Business Intelligence com o Microsoft SQL Server 2005 – Parte 1**. Rio de Janeiro: SQL Magazine, 2009.

ROCHA, Reydeval. **Introdução ao Analysis Services 2005 – Parte 2**. Disponível em:.<http://www.devmedia.com.br/articles/post-5730-Introducao-ao-Analysis-Services-2005-Parte-2.html>. Acesso em: 26 out. 2010.

SINGH, Harry S. **Data warehouse: Conceitos, Tecnologias, Implementação e Gerenciamento**. São Paulo: Makron Books, 2001.

SANTOS, Ricardo S.; ALMEIDA, André Luiz; TACHINARDI, Umberto; GUTIERREZ, Marco Antônio. **Data Warehouse para a Saúde Pública: Estudo de Caso SES-SP**. Disponível em: <http://www.incor.usp.br/spdweb/prodcient_eng/ filesTrabalhos2006/CBIS2006RS.pdf> Acesso em: 14 mar. 2014.

TALEND. **Data Integration**. Disponível em: <http://www.talend.com/ products/data-integration> Acesso em: 25 fev. 2014.

TAVARES, Edmir de Jesus Oliveira. **Processo ETL: O Caso da Unitel T+ Telecomunicações**. Disponível em: <http://bdigital.cv.unipiaget.org:8080/jspui/bitstream/10964/504/1/Memo_Edmir_Tavares%20ESI-SI.pdf> Acesso em: 25 fev. 2014.

WITHEE, Ken. **Microsoft Business Intelligence for Dummies**. Hoboken: Wiley Publishing, Inc., 2010.

APÊNDICES

Apência A – Código SQL Data Warehouse

```
-- PAC_ESTAGIO

CREATE TABLE `PAC_ESTAGIO` (

  `IDN_ESTAGIO`         INT(11) NOT NULL,

  `ESTAGIO`             TINYTEXT,

  PRIMARY KEY (`IDN_ESTAGIO`));


-- PAC_DIGS

CREATE TABLE `PAC_DIGS` (

  `IDN_DIGS`    INT(11) NOT NULL,

  `SUBEIXO`     VARCHAR(21),

  `TIPO`        VARCHAR(38),

  PRIMARY KEY (`IDN_DIGS`));


-- PAC_EXECUTOR

CREATE TABLE `PAC_EXECUTOR` (

  `IDN_EXECUTOR`     INT(11) NOT NULL,

  `TXT_EXECUTORES`VARCHAR(200),

  PRIMARY KEY (`IDN_EXECUTOR`));
```

```
-- PAC_LOCAL

CREATE TABLE `PAC_LOCAL` (

  `IDN_EMPREENDIMENTO`    INT(11) NOT NULL,

  `SIG_UF`                VARCHAR(2),

  `TXT_MUNICIPIOS`        VARCHAR(300),

  `VAL_LAT`               VARCHAR(20),

  `VAL_LONG`              VARCHAR(20),

  PRIMARY KEY (`IDN_EMPREENDIMENTO`));


-- PAC_ORGAO

CREATE TABLE `PAC_ORGAO` (

  `IDN_ORGAO`         INT(11) NOT NULL,

  `DSC_ORGAO`         VARCHAR(50),

  PRIMARY KEY (`IDN_ORGAO`));


-- PAC_TEMPO

CREATE TABLE `PAC_TEMPO` (

  `IDN_TEMPO`         INT(11) NOT NULL,

  `DIA`               DATE,

  `MES`               VARCHAR(9),

  `ANO`               INT(4),

  `NROMES`            INT(2),

  `NROMESANO`         INT(6),
```

```
  `NROANOMES`          INT(6),

  `TRIMESTRE`          INT(1),

  `SEMESTRE`           INT(1),

  PRIMARY KEY (`IDN_TEMPO`));


-- PAC_EMPREENDIMENTO

CREATE TABLE `PAC_EMPREENDIMENTO` (

  `IDN_EMPREENDIMENTO`   INT(11) NOT NULL,

  `DAT_CICLO`            DATE,

  `VAL_2011_2014`        DOUBLE,

  `VAL_POS_2014`         TINYTEXT,

  `INVESTIMENTO_TOTAL`   TINYTEXT,

  `EMBLEMATICA`          TINYTEXT,

  `IDN_EXECUTOR`         INT(11),

  `IDN_ORGAO`            INT(11),

  `IDN_ESTAGIO`          INT(11),

  `IDN_DIGS`             INT(11),

  PRIMARY KEY (`IDN_EMPREENDIMENTO`),

  KEY `IDN_ESTAGIO` (`IDN_ESTAGIO`),

  KEY `IDN_DIGS` (`IDN_DIGS`),

  KEY `IDN_EXECUTOR` (`IDN_EXECUTOR`),

  KEY `IDN_ORGAO` (`IDN_ORGAO`),

  CONSTRAINT `FK_LOCAL` FOREIGN KEY (`IDN_EMPREENDIMENTO`) REFERENCES
`PAC_LOCAL` (`IDN_EMPREENDIMENTO`),
```

```
 CONSTRAINT `FK_ESTAGIO` FOREIGN KEY (`IDN_ESTAGIO`) REFERENCES `PAC_ESTAGIO` (`IDN_ESTAGIO`),

 CONSTRAINT `FK_DIGS` FOREIGN KEY (`IDN_DIGS`) REFERENCES `PAC_DIGS` (`IDN_DIGS`),

 CONSTRAINT `FK_EXECUTOR` FOREIGN KEY (`IDN_EXECUTOR`) REFERENCES `PAC_EXECUTOR` (`IDN_EXECUTOR`),

 CONSTRAINT `FK_ORGAO` FOREIGN KEY (`IDN_ORGAO`) REFERENCES `PAC_ORGAO` (`IDN_ORGAO`));
```

Apência B – Código SQL Estrutura Original

```
-- PAC

CREATE TABLE PAC (

IDN_EMPREENDIMENTO          INTEGER,

IDN_DIGS                    INTEGER,

DSC_TITULO                  VARCHAR(200),

VAL_2011_2014               INTEGER,

VAL_POS_2014                INTEGER,

INVESTIMENTO_TOTAL          INTEGER,

SIG_UF                      VARCHAR(50),

TXT_MUNICIPIOS              VARCHAR(400),

TXT_EXECUTORES              VARCHAR(200),

DSC_ORGAO                   VARCHAR(200),

IDN_ESTAGIO                 INTEGER,

DAT_CICLO                   DATE,

DAT_SELECAO                 DATE,

DAT_CONCLUSAO_REVISADA      DATE,
```

```
VAL_LAT                         VARCHAR(20),

VAT_LONG                        VARCHAR(20),

EMBLEMATICA                     VARCHAR(20),

OBSERVACAO                      VARCHAR(200));


-- DIGS

CREATE TABLE DIGS (

IDN_DIGS        INTEGER,

SUBEIXO         VARCHAR(50),

TIPO            VARCHAR(50));


-- ESTAGIO

CREATE TABLE ESTAGIO (

IDN_ESTAGIO  INTEGER,

ESTAGIO         VARCHAR(50),

DESCRICAO    VARCHAR(200));
```