

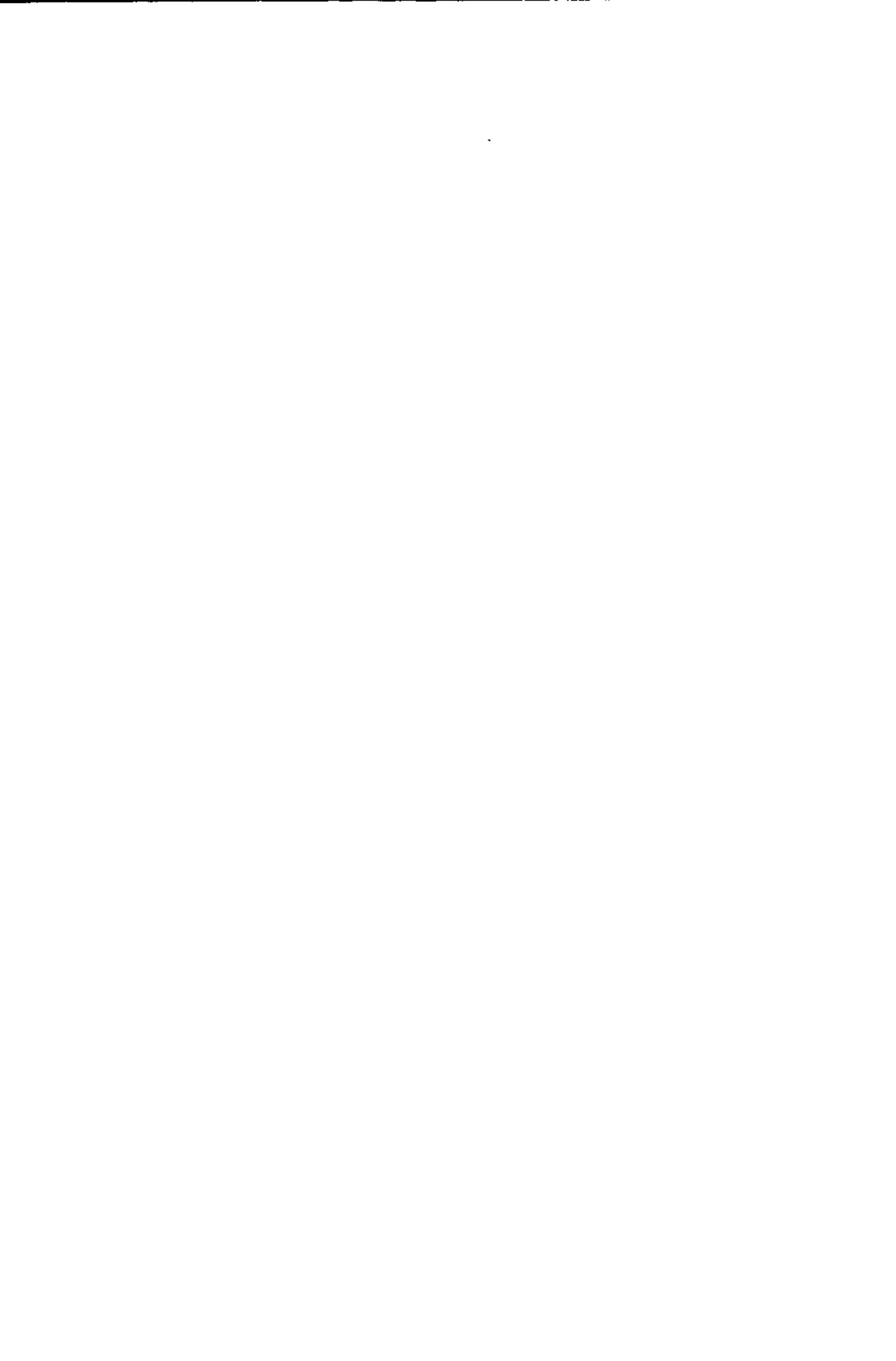
UNIVERSIDADE ESTADUAL DO PIAUÍ – UESPI
CAMPUS PROF. ALEXANDRE ALVES DE OLIVEIRA
BACHARELADO EM CIÊNCIAS DA COMPUTAÇÃO

CONRADO DE SAMPAIO MACHADO NETO

USO DE UM DATA WAREHOUSE TENDO COMO ESTUDO DE CASO
A EMPRESA MAGAZINE ROSEMARY LTDA

PARNAÍBA – PI

2010



CONRADO DE SAMPAIO MACHADO NETO

**USO DE UM DATA WAREHOUSE TENDO COMO ESTUDO DE CASO
A EMPRESA MAGAZINE ROSEMARY LTDA**

Monografia apresentada à Coordenação do Curso de Bacharelado em Ciências da Computação da Universidade Estadual do Piauí – UESPI, Campus Prof. Alexandre Alves de Oliveira, como requisito parcial para a obtenção do título de Bacharel em Ciências da Computação.

Orientador (a): Alessandro Saraiva Mendonça

FICHA CATALOGRÁFICA ELABORADA PELO BIBLIOTECÁRIO
HERNANDES ANDRADE SILVA CRB-3/936

M149u Machado Neto, Conrado de Sampaio

1. Uso de um Data Warehouse tendo como estudo de caso a empresa Magazine Rosemary Ltda / Conrado de Sampaio Machado Neto. – Parnaíba: 2010.
* 60f: il.

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação, Universidade Estadual do Piauí - UESPI, Parnaíba - 2010.

Orientador: Prof. Alessandro Saraiva Mendonça.

1. Banco de Dados. 2. Data Warehouse. 3. Arquitetura de Computação. I. Título.

CDD – 005.756 5



Ata de Apresentação de Trabalho de Conclusão de Curso

Aos dez dias do mês de setembro de dois mil e dez, às 08h00, na Sala 202 do Campus Prof. Alexandre Alves Oliveira - UESPI, na presença da Banca Examinadora, presidida pelo Prof. Alessandro Saraiva Mendonça e composta pelos membros efetivos os professores Francisco das Chagas Rocha e José Flávio Gomes Barros, o aluno **Conrado de Sampaio Machado Neto** apresentou o 'Trabalho de Conclusão de Curso intitulado **Uso de um Data Warehouse Tendo como Estudo de Caso a Empresa Magazine Rosemary Ltda**, como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação. Aberta a sessão pública, o candidato teve a oportunidade de expor o trabalho. Após a exposição, o aluno foi arguido oralmente e avaliado em sessão reservada pelos membros da Banca, nos termos do Regulamento Geral dos Trabalhos de Conclusão de Curso da Universidade Estadual do Piauí, tendo obtido nota 7,3 (sete pontos e três décimos). A Banca concluiu pela **aprovação do candidato, sem restrições**, resultado este divulgado ao aluno e demais presentes. Nada mais havendo a tratar, eu, Prof. Alessandro Saraiva Mendonça lavrei a presente ata que será lida e assinada por mim e pelos membros da Banca Examinadora e pelo candidato. Parnaíba(PI), 10 de setembro de 2010.

Banca Examinadora

Alessandro Saraiva Mendonça
Prof. Esp Alessandro Saraiva Mendonça, UESPI
Presidente

Francisco das Chagas Rocha
Prof. MsC Francisco das Chagas Rocha, UESPI

José Flávio G. Barros
Prof. Esp José Flávio Gomes Barros, UESPI

Candidato

Conrado de Sampaio Machado Neto
Conrado de Sampaio Machado Neto

AGRADECIMENTOS

- Agradeço por todos que aplicaram não só confiança, mas também forças para que eu continuasse seguindo e alcançasse meus objetivos;
- Aos meus pais Paulo Rodrigues e Liduína Machado pelo amor, carinho, apoio exemplo de vida e de garra;
- Meus irmãos Paulo, Jannaina e Larissa pelo amor, carinho e apoio;
- Minha avó Terezinha pela ajuda e apoio, não esquecendo meus tios e primos por sempre estarem presentes na minha vida;
- À Rodrigues, minha segunda mãe, pelo força, proteção e ajuda;
- Aos amigos que conquistei em Parnaíba e na Casa Universitária pela amizade;
- À Tatiane, amiga que me ajudou bastante nessa luta e Layanne pela paciência, contribuição e carinho;
- Aos amigos de faculdade, em especial, Marcel Moura, Alessandra Braúna e Anderson Passos, por toda contribuição, ajuda, compreensão e amizade verdadeira;
- A todos os professores que por esses quatro anos de luta participaram e influenciaram diretamente no meu conhecimento e sem esquecer a Universidade Estadual do Piauí.
- Obrigado a todos, de Coração!

Dedico este trabalho a priori à minha família e a todos os meus amigos que me ajudaram nessa luta.

A perfeição é alcançada não quando não há mais o que acrescentar, mas sim quando não há mais o que tirar.

(Antoine de Saint-Exupéry)

RESUMO

Quase todas as empresas necessitam hoje de um Banco de Dados para armazenar, além dos dados, tão importantes para elas, as ferramentas de acesso fazendo-o um agente facilitador no processo de realização de transações. O problema é que hoje apenas o Banco de Dados não é o bastante, porque primeiro suas ferramentas possuem uma enorme deficiência para analisar uma grande quantidade de dados nele contido, até mesmo por esses dados talvez não estarem concentrados apenas em uma máquina. Para que as empresas possam continuar competitivas, pretendendo maximizar os seus lucros e reduzir possibilidades de erros no processo de tomada de decisão, foi necessário o desenvolvimento de novas ferramentas que auxiliam nesse processo. Uma delas foi o Data Warehouse, que em conjunto com outras ferramentas de bastante valia como Data Mart, Data Mining, realizam trabalhos de análise e tomada de decisões tão importantes para as empresas continuarem dentro do mercado. A partir dessas novas tecnologias, surgiu o conceito de Business Intelligence. Esse estudo foi direcionado para o estudo de caso de um Data Warehouse para a empresa Magazine Rosemary LTDA, trabalho esse que pode auxiliar na maximização dos lucros, organização interna e possibilidade de geração de relatórios constantes dos dados e logo, gerenciamento eficaz de informação e tratamento dos dados nele contidos, para que assim possam ser tomadas boas decisões e com menor probabilidade de erro.

Palavras-chave: Banco de Dados. Data Warehouse. Data Mart. Data Mining. Business Intelligence.

ABSTRACT

Nowadays almost all businesses companies need a Database to store their information, besides the data which is so important for them, access keys make it a facilitator in the process of completing transactions. The trouble is that today the Database only is not enough, because first that your tools have a huge failure to analyze a large amount of data by itself, even this data may not be concentrated on just one machine. For companies to carry on being competitive, aiming to maximize their profits and reduce the chances of errors in decision-making process it was necessary to develop new tools that will assist in this process. One was the Data Warehouse, which set with other tools quite useful as Data Mart, Data Mining, held for analysis and decision's so important allowing companies to continue trading. Based on these new technologies they came up the concept of Business Intelligence. This study was directed by the research case of a Data Warehouse for the enterprise Magazine Rosemary LTDA, this work may maximize profits, internal organization and the possibility of

Keywords: Data Base. Data Warehouse. Data Mart. Data Mining. Business Intelligence.

LISTA DE GRÁFICOS

Gráfico 01: Quantidade de vendas realizadas no período de 02/01/2009 a 05/01/2009 por cada vendedor.	49
Gráfico 02: Quantidade de vendas realizadas por cada funcionário no ano de 2009.	51
Gráfico 03: Quantidade de vendas realizadas pelas três funcionárias com maior número de vendas no ano de 2009.	52

LISTA DE TABELAS

- Tabela 01:** Quantidade de vendas realizadas por cada funcionário no período de 02/01/2009 a 05/01/2009, ordenadas de forma decrescente em relação à quantidade de vendas. 50
- Tabela 02:** Quantidade de vendas realizadas por todos os funcionários durante o ano de 2009, ordenadas de forma decrescente em relação à quantidade de vendas. 51
- Tabela 03:** Quantidade de vendas realizadas pelas três funcionárias com maior número de vendas durante o ano de 2009, ordenadas de forma decrescente em relação à quantidade de vendas. 52

LISTA DE FIGURAS

Figura 01: Estrutura de um Metadado; Fonte.	28
Figura 02: Estrutura de um MDB com duas dimensões.	30
Figura 03: Estrutura de um MDB com três dimensões.	30
Figura 04: Data Mart Independentes.	36
Figura 05: Data Marts Independentes em um Data Warehouse Empresarial.	37
Figura 06: Abordagem Top-Down.	40
Figura 07: Abordagem Bottom-Up.	42
Figura 08: Abordagem Federada.	43
Figura 09: Tela do Sistema operando manualmente.	46
Figura 10: Sistema operando automaticamente.	47
Figura 11: Sistema realizando consultas pré-definidas no Fato Vendas.	47
Figura 12: Sistema realizando consultas não definidas no Fato Vendas.	48

LISTA DE SIGLAS E ABREVIATURAS

DB – Data Base (Banco de Dados)

DDS – Dimensional Data Store (Armazenador de dados Dimensional)

DMRT – Data Mart

DSS – Decision Support System (Sistema de Suporte à Decisões)

DW – Data Warehouse

ETL – Extract, Transform and Load (Extrair, Transformar e carregar)

HOLAP – Hybrid On Line Analytical Processing

KDD – Knowledge Discovery In Data Base

MDB – Multidimensional Data Base

MER – Modelo Entidade Relacionamento

MOLAP – Multidimensional On Line Analytical Processing

NDS – Normalized Data Store

ODS – Operational Data Store (Sistema de Armazenamento Operacional)

OLAP – On Line Analytical Processing

OLTP – On Line Transaction Processing

ROI – Investment and Return On Investment (Investimento de Longo Prazo e Retorno sobre Investimento)

ROLAP – Relational On Line Analytical Processing

SGBD – Sistema de Gerenciamento de Banco de Dados

SUMÁRIO

1 INTRODUÇÃO	13
2 CONCEITOS DE DATA WAREHOUSE	15
2.1 HISTÓRICO EVOLUTIVO	16
2.1.1 Sistemas de Gerenciamento de Banco de Dados	17
2.1.2 Data Warehouse	18
2.2 FILOSOFIAS DO DATA WAREHOUSE	18
2.2.1 Dados empresariais	19
2.2.2 Assunto-Orientação	19
2.2.2 Integração	20
2.2.2 Dados não voláteis	21
2.2.2 Tempo Variante	21
2.2.2 Versão Única da Verdade	22
2.2.2 Investimento de Longo Prazo e Retorno sobre Investimento (ROI)	22
2.3 ARQUITETURAS DE DATA WAREHOUSE	23
2.4 METADADOS	28
2.5 BANCO DE DADOS MULTIDIMENSIONAL	29
2.6 FERRAMENTAS DE DATA WAREHOUSE	31
2.6.1 Data Mining	31
2.6.2 On Line Transaction Processing	32
2.6.3 On Line Analytical Processing	34
2.6.3.1 Multidimensional On Line Analytical Processing – MOLAP	34
2.6.3.2 Relational On Line Analytical Processing – ROLAP	35
2.6.3.3 Hybrid On Line Analytical Processing – HOLAP	35
2.7 DATA MART	35
2.7.1 Data Mart versus Data Warehouse	38
2.7.2 Tipos de Abordagens	39
2.7.2.1 Abordagem Top-Down	39
2.7.2.2 Abordagem Bottom-Up	41
2.7.2.3 Abordagem Federada	43
3 DESENVOLVIMENTO E OBTENÇÃO DE RESULTADOS	45
3.1 A EMPRESA MAGAZINE ROSEMARY LTDA.	45
3.2 CONHECENDO O SISTEMA E AS TECNOLOGIAS UTILIZADAS	45
3.3 CONSTRUÇÃO DO DW	48
3.3.1 Demonstrações	49
4 CONSIDERAÇÕES FINAIS	53
REFERÊNCIAS BIBLIOGRÁFICAS	54
Anexo A: Tabelas utilizadas para montar o cubo “Fato_Vendas”	58
Anexo B: Esquema Snowflake do cubo “Fato_Vendas”	59
Anexo C: Fluxograma de Construção de um Data Warehouse	60

1 INTRODUÇÃO

Historicamente, a sociedade teve a necessidade de armazenar informações de todos os tipos, que se fazem necessárias para confirmação de dados, evitar perda de informações valiosas e etc. O meio utilizado antes dos grandes bancos de dados eram os arquivos impressos, tão úteis que mesmo com o dado digital armazenado, hoje ainda muitas empresas preservam seus arquivos documentados em papéis.

A partir da década de 90 – com o grande crescimento do uso de computadores pessoais – o avanço da internet propiciou a facilidade de acesso a dados em qualquer lugar. Muitas empresas começaram a perceber que poderiam agilizar esse processo de acesso aos dados otimizando processos de consulta, fazendo assim um diferencial na competição entre as outras. Havendo a necessidade de agilizar esse processo de busca, desenvolvedores tiveram a necessidade de criar novas estruturas que suportassem tal plataforma através de conceitos de multidimensionalidade e navegabilidade hierárquica. A partir desse desenvolvimento, deu-se o conceito de Business Intelligence que oferece serviços de extração de metadados potencialmente organizados, explorando assim o processo de Data Warehousing.

Estudar e conhecer um Data Warehouse pode ajudar os gestores a descobrir novas formas de competir em uma economia globalizada e competitiva, trazendo melhores produtos ou serviços para o mercado, mais rápido do que os concorrentes, sem aumentar o custo do produto ou do serviço. Não existem ainda metodologias formais para implementação de um Data Warehouse. Ela deve ser adaptada às características, às expectativas e necessidades de cada empresa, mas o principal objetivo em todas elas é o de descobrir maneiras diferentes de atuar no mercado e quais mudanças internas que devem ocorrer para atender as novas realidades.

Um Data Warehouse não representa nada sem suas poderosas ferramentas de exploração de dados, repositórios especializados e também ferramentas que operam com busca de conhecimento. Essas ferramentas tratam-se de Data Mining, que se baseia na busca de anomalias e possíveis relacionamentos, descobrindo os problemas e oportunidades escondidas nos relacionamentos, diagnosticando o comportamento do negócio e requerendo a mínima intervenção do usuário, Data

Mart, considerados repositórios de dados orientados por assunto, ferramentas de busca OLAP e etc.

No próximo capítulo, segue um breve histórico sobre os primeiros bancos de dados, sua forma de utilização e evolução, explicando também a criação de suas principais ferramentas para organização, análise e tomada de decisões bem como seu valor nos dias de hoje, não só para empresas, mas também em todas as áreas em que essas ferramentas são potencialmente úteis e utilizadas.

Seguindo, trabalhamos o Data Warehouse em si, seus princípios, metodologias, arquiteturas, tecnologias necessárias para realizar um Data Warehousing, junto com conceitos relacionados a metadado, Mineração de Dados, Data Mart e ferramentas de busca.

Por fim, no último capítulo, demonstramos o funcionamento das tecnologias utilizadas, da ferramenta criada para o processo de ETL¹, o processo de construção do Data Warehouse em si e os resultados obtidos realizando operações diretamente do Data Warehouse.

Para concluir este trabalho, referencio a cada ferramenta utilizada, bem como das que podem ser utilizadas para obter melhores resultados. Cito também como trabalho futuro a migração para uma plataforma livre.

¹ Extract, Transform and Load (Extrair, Transformar e Carregar)

2 CONCEITOS DE DATA WAREHOUSE

Vários autores discernem seus próprios conceitos a respeito de Data Warehouse, pois não há um conceito único sobre o assunto.

Data Warehouse é um armazém de dados, ou seja, um conjunto de dados integrados de apoio a decisões gerenciais, e tem como um objetivo satisfazer as necessidades dos vários usuários pelo armazenamento dos dados, que servem para fazer consultas bem como uma visão clara dos negócios que estão acontecendo na empresa (INMON, p. 29, 2005).

Um Data Warehouse é um recurso de uma empresa e existe para o benefício da empresa inteira. Ele não existe para o benefício de uma única entidade dessa empresa (por exemplo, unidade de negócios, o cliente em individual, etc). Em vez disso, um Data Warehouse destina-se a fornecer dados para a empresa inteira, de tal maneira que todos os membros podem usar os dados no Data Warehouse ao longo de sua existência. (SILVERS, p. 11, 2008).

Basicamente, um DW é uma coleção de dados com determinadas características. Esta coleção de dados é normalmente utilizada para fins de análise de propósitos, está orientada a sujeitos, variantes no tempo, integrados e limpos para se acomodar a um padrão de entendimento ou definição de conteúdo e significado, e otimizado para suportar a análise DSS (Sistema de Suporte à Decisões) a ferramentas de Business Intelligence e as funções de uma organização. A maioria dos Data Warehouses contém grandes quantidades de dados históricos, que variam de três para dez anos.

O desenvolvimento dos Data Warehouses segue uma metodologia própria, distinta da que é normalmente aplicada ao projeto de bancos de dados para fins transacionais. Seu objetivo é casar as expectativas dos usuários, em termos de análise e tomada de decisão, com os dados disponíveis. A modelagem adotada prioriza o desempenho nas consultas, adotando modelos mais simples, a chamada modelagem dimensional, ou esquema-estrela. Sua construção inclui, ainda, a limpeza de dados, que leva a um estado mais consistente dos mesmos que podem ser para tomada de decisão.

Segundo ELMARSI e NAVATHE (2000), grandes institutos de pesquisas, tais como, International Data Corporation, IDC, Gartner Group e outros, dizem que no

futuro, Data Warehouse e Data Mining serão bastante explorados no mundo da informática (hoje em dia é a de gestão empresarial), em um mercado que movimentará cerca de US\$ 2 bilhões. E não é para menos, pois este tipo de trabalho é extremamente estimulante para várias, ou todas as áreas de uma empresa.

Um Data Warehouse armazena dados analíticos, destinados às necessidades da gerência no processo de tomada de decisões. Isto pode envolver consultas complexas que necessitam acessar um grande número de registros, por isso é importante a existência de muitos índices criados para acessar as informações da maneira mais rápida possível. Ele armazena informações históricas de vários anos e, por isso, deve ter uma grande capacidade de processamento e armazenamento de dados que se encontram de duas maneiras, detalhados e resumidos.

No Data Warehouse os dados podem ser retirados de múltiplos sistemas de computação normalmente utilizados há vários anos e que continuam em operação, como também podem ser de fontes externas da empresa. Data Warehouses são construídos para que tais dados possam ser armazenados e acessados de forma que não sejam limitados por tabelas e linhas estritamente relacionais. Os dados de um Data Warehouse podem ser compostos por um ou mais sistemas distintos e sempre estarão separados de qualquer outro sistema transacional, ou seja, deve existir um local físico onde os dados desses sistemas serão armazenados.

2.1 HISTÓRICO EVOLUTIVO

Pode-se dizer que tudo começou na década de 60, nos computadores baseados em transistores que, com o passar do tempo, tomaram-se mais acessíveis e empresas passaram a utilizar ferramentas computacionais para processamento de dados internos. Mesmo já com a utilização de computadores, o armazenamento de informações era feito por meios físicos, ou seja, em pastas, grandes arquivos que consumiam muito espaço em salas e eram de difícil locomoção. Sendo assim, em uma empresa, cada filial teria que ter sua própria sala de arquivos, correndo assim risco de perdas parciais ou total de informações. Quando surgiu a idéia de armazenamento de dados, ela era utilizada apenas no momento do processo de gravação e geração de relatórios.

Os sistemas de processamentos de alto desempenho trouxeram várias facilidades, seu uso em empresas cresceu exponencialmente, com isso, o uso de sistemas de armazenamento de dados. Logo após o uso dos cartões magnéticos, veio o uso das fitas DAT (fitas magnéticas), onde houve o crescimento do armazenamento de dados, e junto problemas como dados duplicados e/ou redundantes, pois como se sabe, o armazenamento era feito de forma sequencial, onde para utilizar pelo menos uma parte do sistema, teria primeiro que carregar todo o sistema para poder usar o que realmente era necessário.

2.1.1 Sistemas de Gerenciamento de Banco de Dados

Em busca de soluções para os problemas que existiam, em 1961, foi criado o primeiro Data Base Manager System – Sistema Gerenciador de Banco de Dados (SGDB). Em conjunto com a General Electric, Charles Bachman desenvolveu um dos primeiros sistemas de gerenciamento de banco de dados, denominado de Integrated Data Store (IDS), onde ficava encarregado em gerenciar dados inseridos em um banco. Bachman também trabalhou nos diagramas de estrutura de dados. (ELMARSÍ e NAVATHE, 2000.)

Um SGDB é, portanto, um conjunto de dados associados a um conjunto de programas de acesso a esses dados. Seu principal objetivo é proporcionar um ambiente tanto conveniente quanto eficiente para recuperação e armazenamento de informações.

Um SGDB é projetado para administrar e gerenciar uma grande massa de informações, onde esse gerenciamento implica em definição de estruturas de armazenamento dessas informações e em definição de mecanismos para a manipulação das próprias (SILBERSCHARTZ, P. 17. 1999).

Além do mais, o sistema de gerenciamento deve garantir a segurança das informações que ele administra contra eventuais problemas no sistema, além de impedir tentativas de acesso não autorizado. Essa informação, considerada o que há de valor em uma empresa, tem determinado o estudo para o desenvolvimento de ferramentas cada vez mais eficazes e seguras para diversos tipos de manipulação de dados.

Uma coleção de dados operacionais inter-relacionados. Estes dados são armazenados de forma independente dos programas que os utilizem, servindo a múltiplas aplicações de uma organização (SILBERSCHARTZ, 1999, p. 01).

Portanto, um SGDB é um sistema de software de propósito geral que facilita os processos de definição, construção, manipulação e compartilhamento de bancos de dados entre vários usuários e aplicações, além de realizar processos triviais de inserção, remoção e atualização de dados.

Um SGDB não possui apenas um banco de dados, mas também uma completa definição e descrição da estrutura desse banco. Essa definição está localizada no catálogo do SGDB, onde contém informações como a estrutura de cada arquivo, o tipo de dado e seu devido formato armazenado em cada item. Essas informações armazenadas nesse catálogo são chamadas de metadados e definem a estrutura do banco de dados primário (ELMARSI e NAVATHE, p. 07. 2000).

2.1.2 - Data Warehouse

Na década de 1990, Kimball e Inmon criaram e documentaram os Conceitos e Princípios de Data Warehouse. Estes conceitos não dirão ao desenvolvedor como construir um Data Warehouse, mas auxiliam-no a compreender quais os motivos para criação. Por isso, esses conceitos e princípios são conhecidos coletivamente como "Filosofia dos Data Warehouses".

Os Data Warehouses surgiram como conceito acadêmico na década de 1980. Com o amadurecimento dos sistemas de informação empresariais, as necessidades de análise dos dados cresceram paralelamente. Os sistemas de processamento de transações, OLTP, não conseguiam cumprir a tarefa de análise com a simples geração de relatórios. Nesse contexto, a implementação do Data Warehouse passou a se tornar realidade nas grandes corporações. O mercado de ferramentas de Data Warehouse, que faz parte do mercado de Business Intelligence² cresceu então, e ferramentas melhores e mais sofisticadas foram desenvolvidas para apoiar a estrutura do Data Warehouse e sua utilização.

2.2 FILOSOFIAS DO DATA WAREHOUSE

Para a criação e conseqüentemente o sucesso de um Data Warehouse, não há necessidade de seguir todos os princípios inclusos na Filosofia. O desenvolvedor ter a consciência dos princípios utilizados aumentam o valor e o sucesso do projeto. O desenvolvedor do Data Warehouse pode optar por incluir ou excluir elementos

² Serviços de extração de dados potencialmente organizados para fins empresariais

contidos na Filosofia. Essas decisões devem ser feitas de acordo com o contexto utilizado no momento da criação. Esses elementos são:

- Dados empresariais (Enterprise Data)
- Assunto-Orientação (Subject Orientation)
- Integração (Integration)
- Dados não voláteis (Non-volatility)
- Tempo Variante (Time Variant)
- Versão Única da Verdade (Single Version of the Truth)
- Investimento de Longo Prazo e Retorno sobre Investimento (Long-Term Investment and Return on Investment (ROI))

2.2.1 Dados empresariais

Um Data Warehouse deverá conter dados aplicáveis à empresa. O valor e a relevância de um Data Warehouse está diretamente ligado a seus dados. Se os membros da empresa localizam dados supérfluos e/ou irrelevantes, os mesmos tratarão esses dados para uma nova percepção. Podemos citar como exemplos:

- O Departamento de Contabilidade de uma empresa utiliza códigos fiscais. Membros de outros departamentos entendem a importância daqueles códigos para a empresa.
- O Departamento de Produção utiliza número de peças. Membros de outros departamentos entendem a importância daqueles códigos para a empresa.

Códigos fiscais não são diretamente utilizados no Departamento de Produção, bem como os números de peças não são ao Departamento de Contabilidade, mas os dois departamentos compreendem a relevância daqueles dados para seu determinado departamento.

2.2.2 Assunto-Orientação

Os dados em um Data Warehouse são organizados em torno de assuntos e negócios da empresa. Os dados operacionais são organizados pelas suas manifestações físicas, incluindo os nomes de arquivos, agendas de trabalho e as dependências do aplicativo. Um DW não apresenta os dados diretamente ao

usuário, e sim os apresenta refletindo as principais áreas requisitadas dentro da empresa. Por exemplo:

- Entidades Empresariais
 - Clientes
 - Vendedores
 - Agentes

- Processos Empresariais
 - Vendas
 - Receber
 - Fabricação
 - Distribuição

O elemento Assurto-Orientação permite que o DW mantenha sua arquitetura intacta ao longo de seu tempo de vida. Esse elemento admite que o DW absorva as mudanças inevitáveis, sem alterações drásticas na sua arquitetura.

2.2.3 Integração

Os dados em um DW são exibidos de uma maneira uniforme. Ao integrar os dados, um DW apresenta aos seus clientes uma informação concisa e contínua da empresa, que os alivia da necessidade de conciliar as diferenças e inconsistências nos dados nas áreas de negócio. Essa integração ocorre de várias maneiras, que podem ser combinadas em três grupos:

- **Formas:** Incluem tipos e layouts de dados. Essa é a maneira em que os dados são apresentados. Exemplos:

Valores monetários podem ser expressados como moeda ou inteiros;

Nomes podem ser expressados como (Last-Name / First-Name) ou (First-Name / Last-Name)

Funções: Incluem o conteúdo e o significado dos dados dentro de cada elemento.

- **Grão:** Refere-se à unidade de medida na qual os dados são expressos. As unidades de negócios podem armazenar dados usando diferentes unidades de medida. Grãos de dados tem duas implicações físicas para um DW.

Primeiro, os dados de grãos despontam informações mais detalhadas, mas a um devido custo. Cada vez que há um aumento relacionado ao detalhe, é consumido mais recurso. Detalhes como, capturar, armazenar e recuperar. Segundo, um DW não pode fornecer dados aos clientes de um grão menos do que o grão em que é armazenado. Data Warehouses devem integrar formas, funções e grãos de diferentes negócios. Uma vez integrados, os dados dos clientes no DW podem percorrer dentro de assuntos de negócios da empresa inteira.

2.2.4 Dados não voláteis

Dados uma vez gravados em um DW, nunca serão excluídos ou atualizados, pois Data Warehouses operam apenas com inserção de dados. As aplicações operacionais manipulam os dados para refletir apenas o estado atual de uma unidade de negócio. Um DW reflete tanto estados históricos como atuais da empresa, apenas inserindo novas linhas. Ele mantém linhas históricas, bem como linhas mais recentes, fazendo assim com que o DW possa apresentar dados no contexto do passado e do presente.

Non-volatility permite que um DW expresse uma empresa ao longo do tempo, mantendo seus dados intactos.

2.2.5 Tempo Variante

Um DW expressa os eventos da empresa ao longo do tempo. Pelo fato dos dados serem não voláteis, um DW pode expressar eventos dentro do contexto histórico. Os dados históricos em um DW fornecem respostas às perguntas dos eventos históricos.

Dados no tempo variante permitem que um DW expresse à empresa a situação em um determinado momento. Um momento no tempo pode ser expresso como um milésimo de segundo, minuto, hora, dia, semana, mês, ano, etc. Apesar dos sistemas de informação capturarem o tempo de forma digital, no contexto digital versus analógico, o tempo é analógico. Cada representação de tempo, portanto, é uma representação digital do tempo analógico.

2.2.6 Versão Única da Verdade

Para todas as perguntas que podem ser respondidas através de dados, uma empresa vai decorrer uma gama de respostas.

O princípio da Versão Única da Verdade permite um DW expressar todo o empreendimento. Quando todos os membros possuem uma única compreensão de uma visão dos elementos de dados, o DW pode ser utilizado como um ponto comum de comunicação da empresa.

2.2.7 Investimento de Longo Prazo e Retorno sobre Investimento (ROI)

Um DW atinge maior ROI (Investment and Return on Investment) através de longevidade e estabilidade. Como o número de áreas integradas em um processo de armazenamento de dados, um Data Warehouse aumenta a expressão da sua empresa. Como a história se acumula de acordo com a variação de tempo em um Data Warehouse, ele aumenta a sua capacidade para responder a questões históricas. Um Data Warehouse deve, portanto, ser arquitetado e desenvolvido como um investimento a longo prazo.

Uma equipe não precisa construir o Data Warehouse por inteiro em um único projeto. O custo seria muito alto e o prazo de entrega seria muito lento. Ao invés disso, um DW deve começar com um ou dois assuntos de negócios (por exemplo: venda, transporte, fabricação, etc.) Então, cada empenho no desenvolvimento posterior do DW acrescenta um outro assunto de negócios, ou um subconjunto de um elemento grande ou complexo.

De acordo com Silvers (2008):

[...] Cada projeto individual no Data Warehouse deve durar entre seis e nove meses. Quando a duração de um projeto individual de Data Warehousing for superior a nove meses de gestão, tipicamente deve-se começar a questionar o ROI do projeto.

Um DW, portanto, é um investimento em longo prazo; na verdade, um Data Warehouse não pode expressar toda a empresa. O sucesso de um Data Warehouse não é a sua capacidade de expressar toda a empresa, mas sim a sua capacidade de retomar os valores para a empresa utilizar nos negócios.

O primeiro projeto do DW de uma empresa define a arquitetura de nível empresarial do Data Warehouse. As decisões tomadas durante o primeiro projeto de

Data Warehouse irão lançar as bases de dados para todos os projetos subsequentes dentro da empresa. Fisicamente, essas decisões irão lançar as bases para as plataformas e infraestruturas que constituirão o Data Warehouse. Pela natureza de longo prazo de um DW, o "retorno" exige muito além do investimento. Se feito corretamente, o investimento deve ser de curta duração, e o retorno deve se estender por anos, senão décadas.

2.3 ARQUITETURAS DE DATA WAREHOUSE

Conforme RAINARDI (2008), existem duas arquiteturas principais para um sistema de Data Warehouse: a primeira é a arquitetura de fluxo, que consiste em um repositório organizado dentro de um Data Warehouse e da forma como os arquivos fluem dos sistemas fonte para o usuário através destes repositórios. A segunda é a arquitetura de sistema que se baseia sobre a configuração física dos servidores, software, Data Warehousing e clientes.

Uma arquitetura de fluxo de dados é uma das primeiras coisas que você precisa para decidir o sistema de Data Warehouse, pois ela determina quais componentes precisam ser construídos e, portanto, afetam o plano do projeto e os custos. A arquitetura de fluxo de dados mostra como os dados fluem através do Data Warehousing em um Data Warehouse.

No uso de Data Warehouse, a arquitetura de fluxo de dados é uma configuração de dispositivos de armazenamento dentro de um Data Warehouse, juntamente com o arranjo da forma como os fluxos de dados dos sistemas de origem através destes Data Stores para as aplicações utilizadas pelos usuários finais. (RAINARDI, 2008, p. 29.)

Como podemos notar é fundamental o uso de Data Stores que formam componentes importantes da arquitetura de fluxo de dados. O armazenamento de dados é oriundo de um ou mais bancos de dados ou arquivos contendo informações de armazenamento, dispostos em um formato especial e envolvidos em processos de Data Warehousing. Sendo que esses dados armazenados em Data Stores são classificados como dados de interface, onde o usuário armazena os dados que estarão disponíveis para usuários finais que em seguida poderão usá-los em suas aplicações. Um Data Warehousing interno é usado pelos componentes de Data Warehouse e possui a finalidade de integrar, limpar e registrar os dados, não estando aberto para consultas pelos usuários finais.

Um processo ETL é um programa que acessa dados de uma série de fontes que preenchem uma tabela alvo. O mesmo consiste de uma série de passos. Segundo RAINARDI (2008), "... um passo é um componente de um processo ETL que não especifica a tarefa." Um exemplo de um passo é uma extração particular de dados de uma fonte de armazenamento de dados onde desempenham certas transformações de dados. Os volumes ETL em um Data Warehouse são gerenciados por um sistema de controle, que é um sistema que gerencia o tempo em que cada volume ETL roda, e provem a capacidade de reiniciar o volume ETL do ponto de erro. O mecanismo que registra cada passo do processo ETL é chamado auditor ETL (ETL audit).

A descrição de cada processo ETL é salvo em um metadado³. Isso inclui a fonte da qual os dados serão extraídos, o alvo para onde será carregado, a informação aplicada, o processo pai e a lista em que cada processo é pré-suposto para rodar. Em um Data Warehouse, os metadados são dispositivos de armazenamento que contêm a descrição da estrutura, dados e processos dentro do Data Warehouse. Isso inclui as definições de dados e mapeamento, a estrutura de dados de cada Data Store, a estrutura dos sistemas fonte, a descrição de cada processo ETL, ou seja, a descrição da quantidade de regras e um registro de todos os processos e atividades em um Data Warehouse.

Os processos de qualidade dos dados são atividades e mecanismos para se certificar que os dados no Data Warehouse foram corretos e completos. Isso é geralmente feito caso necessite checar os dados no seu caminho. Eles também abrigam o mecanismo que informa se o salvamento de dados foi ruim, o que possibilita sua correção. A regra da qualidade de dados é o critério que verifica os dados de um sistema fonte que está dentro do esperado em um formato correto.

A arquitetura de fluxo de dados é projetada com base nos requisitos das aplicações de dados, incluindo os requisitos de qualidade de dados. Aplicações de Data Warehouse requerem dados em diferentes formatos, onde eles ditam os armazenamentos de dados que vão ser preciso. Existem três arquiteturas de fluxo de dados: Single DDS, NDS + DDS e ODS + DDS.

³ Estrutura com conhecimento sobre os dados. Dicionário de dados.

Na arquitetura Single DDS, temos um Data Warehouse dimensional. O DDS consiste então em vários Data Marts. Podemos dizer que o Data Mart dimensional é um grupo de tabelas de fatos relacionados e suas tabelas de dimensão correspondentes, contendo as medições de eventos de negócios que são classificados por suas dimensões. A forma física de um estágio pode ser um banco de dados ou arquivos. O ETL, que extrai dados do sistema, insere a fonte de dados em um banco, ou grava-os como arquivos. Um segundo pacote de ETL pega os dados do estágio, integram-nos ao sistema de origem diferente, aplicam-se algumas regras à qualidade dos dados e coloca os dados consolidados para o DDS. Os dados no DDS também podem ser carregados em bases de dados multidimensionais, tais como, SQL Server Analysis Services e, em seguida, acessados pelos usuários através de aplicações OLAP e mineração de dados.

Uma das vantagens de uma arquitetura Single DDS é que os dados do estágio são carregados diretamente para o Data Warehouse dimensional, sem qualquer tipo de Data Store normalizado antes. A principal desvantagem é que é mais difícil nessa arquitetura, criar um segundo DDS. O DDS em um single DDS na arquitetura é o armazenamento de dados mestre, contendo um conjunto completo de dados em um Data Warehouse, incluindo todas as versões e todos os dados históricos. Às vezes é preciso criar um DDS menor contendo um subconjunto de dados do DDS mestre para efeitos de análise específica.

Na arquitetura de fluxo de dados NDS + DDS, existem três armazenadores de dados: Stage, NDS e DDS. Esta arquitetura é similar à arquitetura Single DDS, mas tem um armazenamento de dados normalizado frente ao DDS. O NDS é na terceira forma normal, relacional ou superior. O objetivo do NDS em primeiro lugar é integrar os dados de vários sistemas de origem; em segundo lugar, ele é capaz de carregar dados em vários DDSs. Ao contrário da arquitetura Single DDS, na arquitetura NDS + DDS pode-se ter vários DDSs.

A principal vantagem desta arquitetura é a facilidade para reconstruir o DDS principal e, além disso, a construção de um novo DDS menor. Isso ocorre porque o NDS é o mestre de armazenamento de dados que contém um conjunto completo deles. Isto lhe permite criar uma partição de dados estáticos e armazenar para fins de análise específica. A segunda vantagem permite, mais facilmente, manter os principais dados em uma repositório de normalização, como o NOS e publicá-lo de lá

2.4 METADADOS

Metadado é um componente bastante importante do ambiente de Data Warehouse (metadado também pode ser chamado de dados sobre dados), no qual também constitui parte do meio de processamento da informação durante o tempo de execução. Mas no mundo dos Data Warehouses, metadados levam a um novo nível de importância, pois permitem a utilização mais eficaz do DW.

Em um Data Warehouse, metadados contêm as definições de dados (o significado e a origem de cada coluna), a definição do Data Warehouse em si (em outras palavras, a estrutura de armazenamento de dados, os processos de ETL e da qualidade dos dados), a definição de respectivos sistemas (por exemplo, a fonte de sistemas) e o uso (que relatórios e cubos são usados por quem e quando) (RAINARD, p. 301, 2008).

Metadados permitem que o usuário final ou analista de DSS navegue através das possibilidades. Por outro lado, quando um usuário começa a trabalhar em um Data Warehouse onde não existe metadados, ele não sabe por onde começar a análise. O usuário deve revistar o Data Warehouse para descobrir se há dados para que não haja desperdício de tempo. Mesmo depois que o usuário verificar, não há garantia de que ele vai encontrar os dados corretos, ou interpretar corretamente os dados encontrados. Com a ajuda de metadados, no entanto, o usuário final pode ir rapidamente para os dados necessários ou determinar o que não existe o que ele busca.

column_key	table_key	column_name	data_type	is_PK	is_FK	is_null	is_identity
...							
126	21	customer_key	int	Y	Y	N	N
127	21	communication_key	int	Y	Y	N	N
128	21	channel_key	int	N	Y	N	N
129	21	subscription_start_date_key	int	Y	Y	N	N
130	21	subscription_end_date_key	int	N	Y	N	N
131	21	subscription_status_key	int	N	Y	N	N
132	21	source_system_code	tinyint	N	N	N	N
133	21	create_datetime	datetime	N	N	N	N
134	21	update_datetime	datetime	N	N	N	N
...							

Figura 01: Estrutura de um Metadado; Fonte: *Building a Data Warehouse With Examples in SQL Server*. 2007.

Metadados agem então como um dicionário para o conteúdo do Data Warehouse. Ficam localizado acima do Data Warehouse e informados de tudo que acontece nele. Tipicamente, metadados armazenam os seguintes itens:

- Estrutura de dados, conhecida pelo programador

- Estrutura de dados conhecida pelo analista de DSS
- Fonte de fornecimento de dados do Data Warehouse
- Transformação de dados que passarão para o Data Warehouse
- Modelo de dados
- Relação entre o modelo de dados e o Data Warehouse
- Histórico de extração de dados

2.5 BANCO DE DADOS MULTIDIMENSIONAL

Um administrador de um Data Warehouse de uma empresa sabe que o objetivo final do Data Warehouse é obter os dados para o usuário e exibí-los em forma de relatórios, e não para a base de dados com intuito apenas de armazenamento como muita gente imagina. Uma base de dados multidimensional é utilizada principalmente para realizar análise e em aplicações em Data Mining.

Nós podemos gastar milhões de dólares e dois anos para a criação de servidores, arquitetando o Data Warehouse, projetar a base de dados, construção dos sistemas ETL, popular os Data Stores, mas se os dados não forem recuperados e apresentados aos usuários, o Data Warehouse é inútil (RAINARD, p. 377. 2008).

De acordo com RAINARD (2008, p. 377):

Um Banco de Dados Multidimensional é uma forma de banco de dados onde os dados são armazenados em células e a posição de cada célula é definida por um número hierárquico, chamado "Dimensão". Cada célula representa um evento e o valor das dimensões mostra quando e onde o evento aconteceu.

Em outras palavras, a estrutura do banco multidimensional armazena e agrega valores, esses valores agregados são pré-computados contendo resumos dos valores de base. Fisicamente, um MDB, Banco de Dados Multidimensional, é um arquivo.

A figura 02 representa uma estrutura de duas dimensões. Nesse exemplo, o MDB se assemelha a uma matriz contendo duas dimensões, cliente e produto. A combinação dessas dimensões gera uma célula que pode conter valores combinados ou então vazio. Cada célula representa um evento e o valor de sua dimensão representa onde e quando aconteceu.

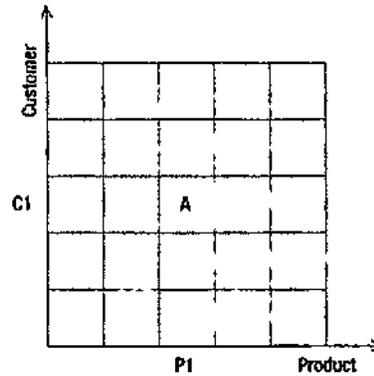


Figura 02: Estrutura de um MDB com duas dimensões;

Fonte: *Building a Data Warehouse With Examples in SQL Server. 2008.*

Já a figura 03, mostra um MDB com três dimensões, com o formato de um cubo. Bases de Dados Multidimensionais com mais de 3 dimensões podem também ser chamadas de Hipercubo. Na figura 3 possui como dimensões, cliente, produto e tempo. Como na estrutura de duas dimensões, a combinação das três dimensões gera uma célula no cubo conhecida como evento. Nesse caso, cada célula mostra o que um determinado cliente comprou naquela faixa de tempo.

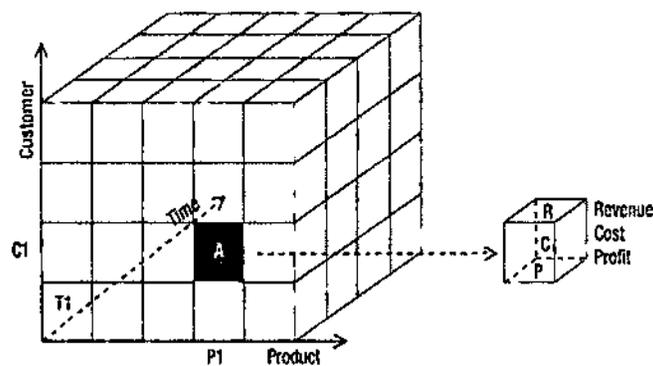


Figura 03: Estrutura de um MDB com três dimensões

Fonte: *Building a Data Warehouse With Examples in SQL Server. 2008.*

Apesar de manter uma organização bem detalhada, a desvantagem de se usar um MDB em comparação a um banco de dados relacional é o tempo de processamento necessário para o carregamento do banco de dados e o cálculo do valor agregado. Sempre que a fonte relacional é atualizada, o MDB precisa ser atualizado ou reprocessado, em outras palavras, as células agregadas precisam ser recalculadas (não tem de ser feito em tempo real). A segunda desvantagem é a escalabilidade: o MDB não pode ser bem dimensionado para um banco de dados muito grande (vários terabytes) ou um grande número de dimensões.

2.6 FERRAMENTAS DE DATA WAREHOUSE

Para um bom funcionamento, e acima de tudo, obter os resultados esperados, foram criadas ferramentas para trabalhar com os Data Warehouses, pois sem elas, um DW é apenas um repositório de dados estruturado. Para fins como buscas por padrões e constantes consultas especialistas criaram Data Mining, OLAP e OLTP, ferramentas hoje imprescindíveis para manipular o Data Warehouse.

2.6.1 Data Mining

Data Mining é um "descobridor" de conhecimentos em base de dados chamado de KDD (Knowledge Discovery In Data Base). Esta ferramenta apresenta uma nova metodologia que melhora a qualidade e influencia as decisões científicas ou de negócios. (FAYYAD, 1996).

Todos os sistemas de Data Warehouse possuem bases de dados bem organizadas e consolidadas que funcionam com boas ferramentas de exploração, por isso a tecnologia do Data Mining é considerada como um minerador de dados. Eles chamaram a atenção dos construtores de Data Warehouse porque a mineração de dados tem alternativas de grandes descobertas de oportunidades de negócios e além disso, também pode traçar estratégias novas para o futuro. (NIMER; KIMBALL, 1998)

Data Mining basicamente é aplicação de técnicas estatísticas, muitas vezes complexas e que precisam ser analisadas por técnicos especializados. O cérebro humano, comprovadamente, consegue fazer até 8 comparações ao mesmo tempo. A função do Data Mining é justamente ampliar esta comparação para "infinito" e tomar isso visível ao olho humano.

Segundo ELMARSI e NAVATHE (2000):

[...] Data Mining pode ser usado em conjunto com o Data Warehouse para auxiliar certos tipos de decisão e também pode ser aplicado a bancos de dados operacionais com transações individuais. Para fazer o Data Mining mais eficiente, o Data Warehouse deve ter uma coleção de dados agregados ou sumarizados. Em relatórios como os do popular Gartner Group, Data Mining tem sido apontada como uma das tecnologias mais promissoras para o futuro próximo.

Para trabalhar com Data Mining, deve-se levar em consideração algumas técnicas que devem ser utilizadas:

- **Análise de Cluster:** Esta técnica agrupa informações homogêneas de grupos heterogêneos entre os demais e aponta o item que melhor representa cada grupo, permitindo, desta forma, que consigamos perceber a característica de cada grupo.
- **Regressão:** As técnicas de regressão auxiliam a previsão do futuro, baseado no passado. Para isso utiliza-se de dados históricos. Existem várias técnicas de regressão e deve-se utilizar a mais adequada em cada caso.
- **Árvore de Decisão:** Nesta técnica o desenvolvedor escolhe a variável que quer avaliar e o software procura as mais correlacionadas e monta a árvore com várias ramificações. As árvores de decisão são meios de representar resultados de Data Mining e lembram um gráfico organizacional horizontal
- **Redes Neurais:** As redes neurais tentam construir representações internas de modelos ou padrões achados nos dados, mas essas representações não são apresentadas para o usuário. Estruturalmente, uma rede neural consiste em um número de elementos interconectados (chamados neurônios) organizados em camadas que aprendem pela modificação da conexão firmemente conectando as camadas.
- **Estatísticas Descritivas:** Para finalizar, as estatísticas descritivas são usadas para, como diz o nome, descrever itens, como o tamanho da população, média de idade e outros mais. Estas técnicas são muito úteis, porém deve haver um técnico que possa se utilizar destas técnicas e de outras mais, como por exemplo: análise de discriminantes, as muitas técnicas de ajustes de curvas, análises de variância, testes de hipóteses, cálculos de probabilidades e outros mais.

2.6.2 On-Line Transaction Processing

Uma aplicação determina o uso de um sistema de processamento de dados em particular, por exemplo, uma aplicação de folha de pagamento ou de uma solicitação de entrada de pedidos. Essas aplicações comerciais geralmente processam muitos pontos semelhantes, por exemplo, uma ordem em um sistema de processamento de pedidos, ou uma consulta de crédito em um sistema de controle

de crédito. O nome dado a cada item de processamento desses citados é transação comercial.

Um importante conceito de On Line Processing Transaction, OLTP, é dado por Humphries (1999, p.9):

O termo online Transaction Processing, OLTP, refere-se a sistemas que automatizam e capturam transações comerciais através da utilização de sistemas de computação. Além disso, estas aplicações tradicionalmente produzem relatórios que permitem aos usuários de negócios controlar o status das transações. Aplicações OLTP e seus respectivos bancos de dados ativos compõem a maioria dos sistemas cliente / servidor de hoje.

Conclui-se que Online Transaction Processing (OLTP) são aplicações cliente-servidor on-line que dão aos usuários acesso direto às informações. O processo de aplicações OLTP, é chamado de transações. Uma única operação pode solicitar uma conta bancária, outra pode atualizar esse equilíbrio para refletir um depósito.

Bancos de dados desenvolvidos para OLTP precisam processar um elevado número de operações (inserir, atualizar e excluir) e ao mesmo tempo recuperando um pequeno conjunto de resultados usados para exibir informações para os usuários do aplicativo. Para isso, o banco de dados tem de ser projetado para ser extremamente eficaz para gerenciar os dados que ele contém.

[...] Bancos de dados com Online Transaction Processing (OLTP) contêm as informações de uso cotidiano das organizações para executar o negócio. Um banco de dados OLTP normalmente contém dados específicos de um processo de negócios e em muitas organizações, você encontrará vários sistemas OLTP funcionando com uma mistura de sistemas operacionais e plataformas de banco de dados. Embora as plataformas que utilizam bases de dados OLTP possam ser diferentes, muitas vezes compartilham o mesmo objetivo de forma que permita aos utilizadores gravar e alterar dados em tempo real. (BAIN et. Al. P. 107. 2001).

Essas bases de dados OLTP devem possuir as seguintes características:

- Suportar um elevado número de utilizadores. Dependendo da natureza da aplicação de banco de dados elas podem suportar uma centena ou milhares de usuários simultâneos.
- Alta taxa de transações. Todos os usuários conectados ao banco de dados não ficarão apenas em espera. Eles vão gerar selecionar, inserir, atualizar e excluir, operações que o banco de dados OLTP deve ser capaz de suportar.
- Integridade dos dados. Com um número tão elevado de usuários gerando um elevado número de transações pode ocorrer erros nos dados o que seria

muito caro para uma organização. Portanto, proteger a integridade dos dados contidos no banco de dados OLTP é de prioridade máxima.

2.6.3 On Line Analytical Processing

A necessidade de receber um grande número de dados de um grande banco de dados (centenas de Giga ou até mais) são os motivos de existir o On Line Analytical Processing, OLAP. Quando temos a necessidade de um sistema multidimensional precisamos de um OLAP.

De acordo com Silvers, (2008, p. 211):

"... OLAP é meramente uma ferramenta analítica. O conjunto de resultados é raramente móvel. O analista usando um aplicativo OLAP deve ter um profundo entendimento do negócio e da empresa para atingir o caminho certo da análise de consciência. O conjunto de resultados e o caminho pelo qual o analista conseguiu o conjunto de resultados geralmente não são repetíveis."

Segundo FIGUEIREDO (1998):

"OLAP são ferramentas que os usuários finais podem explorar e extrair os dados contidos em suas bases, gerando os relatórios de modo satisfatório. Essa forma como os usuários pensam são análises multidimensionais onde os usuários são representados como dimensões e não tabelas".

Uma excelente vantagem do OLAP é que ele é interativo. O analista pode jogar um valor para simular algo. Assim pode inclusive descobrir padrões escondidos.

Para o desenvolvimento de bases multidimensionais, o desenvolvedor pode ter em mãos três tecnologias de OLAP: MOLAP, ROLAP E HOLAP.

2.6.3.1 Multidimensional On Line Analytical Processing – MOLAP

Esta variação de OLAP armazena todos os conjuntos de resultado de todas as permutações de dimensão em um cubo OLAP. MOLAP exige uma boa capacidade de armazenamento. A criação de todos os conjuntos de resultados em um cubo MOLAP exige bons ciclos de CPU e capacidade de memória. MOLAP fornece o desempenho mais rápido para o cliente.

2.6.3.2 Relational On Line Analytical Processing – ROLAP

Já esta, não armazena nenhum conjunto de resultados. Pelo contrário, ROLAP identifica os dados dentro de um Data Warehouse, onde a partir desse pode-se calcular em tempo de execução todos os conjuntos de resultados. Quando um cliente indica um cruzamento de dimensões, o cubo ROLAP traduz essa informação em uma instrução SQL, que é submetido a um Data Warehouse. Um cubo ROLAP exige a menor capacidade de armazenamento no servidor OLAP, no entanto, ROLAP requer mais desempenho direcionado para o Data Warehouse. O desempenho do ROLAP torna-se mais lento e limita o número máximo de dimensões para o cliente.

2.6.3.3 Hybrid On Line Analytical Processing – HOLAP

Por fim, HOLAP, uma combinação de MOLAP e ROLAP que necessita de um pré-cálculo e mais espaço de armazenamento, mas em compensação, nem todos os conjuntos de resultados são armazenados dentro de um cubo OLAP. Um cubo HOLAP alcança um compromisso entre a capacidade, desempenho e permutações de dimensões disponíveis para o cliente.

2.7 DATA MART

O processo de Data Warehousing tomou-se uma arquitetura full-blown⁴, separado dos sistemas operacionais ou transacionais. Hoje, Data Warehousing tem diferentes componentes que servem para diversos fins, dependendo da necessidade. A arquitetura que tem crescido desde as origens do Data Warehouse é algo que hoje pode ser chamada de arquitetura de informações da empresa. Nesta arquitetura, os sistemas operacionais são geralmente On-Line Transaction Processing (OLTP), ou seja, sistemas que atendem a pedidos de usuários finais. Tal aplicação pode ser de ponto de venda, permitindo aos clientes comprar produtos de sua empresa, ou uma aplicação de compra que lhe permite comprar mercadorias de seus fornecedores.

⁴ Totalmente desenvolvido, maduro

Data Marts são considerados depósitos de dados especializados que apoiam os processamentos nos segmentos de um sistema de apoio à decisão. Eles contêm menos dados, são mais baratos e menores que um DW, e são objetos de processo analítico (INMON, 2005; KIMBALL, 1998). Os Data Marts possuem todos os detalhes de dados de um assunto específico, geralmente até mais específico que um Data Warehouse, atendendo várias necessidades de informações. O DW pode conter vários Data Marts, ou seja, sua construção pode se realizar a partir do surgimento de vários Data Marts.

Data Mart é um subconjunto de informações, vindo de um DW desenhado, mantido e distribuído de forma centralizada, alimentando os dados que produzem uma combinação de sistema operacional dos grupos de usuários (INMON, 2005; KIMBALL, 1998).

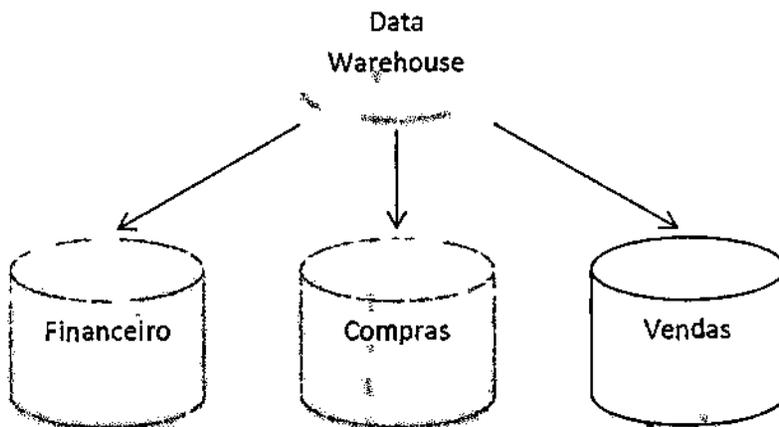


Figura 04: Data Marts Independentes; Fonte: Conrado Machado.

O Processo de Data Mart funciona levando os dados de transações recolhidos através dos sistemas OLTP atravessando uma camada de programas cujo objetivo é integrá-los e transformá-los para que se tomem aptos a entrar no DW. A camada de programas pode ser chamada de "Integração e Transformação". Estes programas podem ser escritos manualmente ou de forma automatizada, através de geradores, programas projetados especificamente para atender às necessidades desta interface.

Depois que os dados transacionais passam através da camada Integração e Transformação, eles são fundamentalmente alterados para atender às necessidades de informações da empresa de acordo com o projeto do DW. As chaves são alteradas – de modo que haja um entendimento corporativo dos dados – as

estruturas são modificadas e os algoritmos de codificação são unificados. Os dados entre a camada de integração e de programas de transformação em um estado considerado "desintegrado" passam para fora da camada de programas em um estado bastante integrado. Os dados, em seguida, passam para o DW da empresa de uma forma muito granular. Uma vez no DW da empresa, os dados estão prontos para servir como base para todos os tipos de sistemas de apoio à decisão (DSS).

O DW também é alimentado pelo ODS. O ODS oferece tempo de resposta padrão de operação (2-3 segundos), bem como serve de local onde os dados podem ser agregados.

Do outro lado do DW, de forma transparente, encontra-se um conjunto de Data Marts. Para o usuário final, um Data Mart aparenta ser um DW, pois é nele que o usuário final tem interação direta com o ambiente de DW. Os Data Marts são orientados para as necessidades e requisitos específicos, tais como os dados que servem as necessidades de um determinado departamento, ou talvez servindo como um subconjunto do DW para um determinado período de tempo, etc. Os Data Marts ou são alimentados de dados granulares do Data Warehouse corporativo, que depois é remodelado para atender as exigências específicas em torno do qual o Data Mart foi arquitetado, ou são alimentados diretamente da camada de integração e transformação. Data Marts, em último caso, fornecem dados ao Data Warehouse com informações que são então mantidas no repositório central da empresa. A figura 05 representa bem a estrutura de um Data Mart.

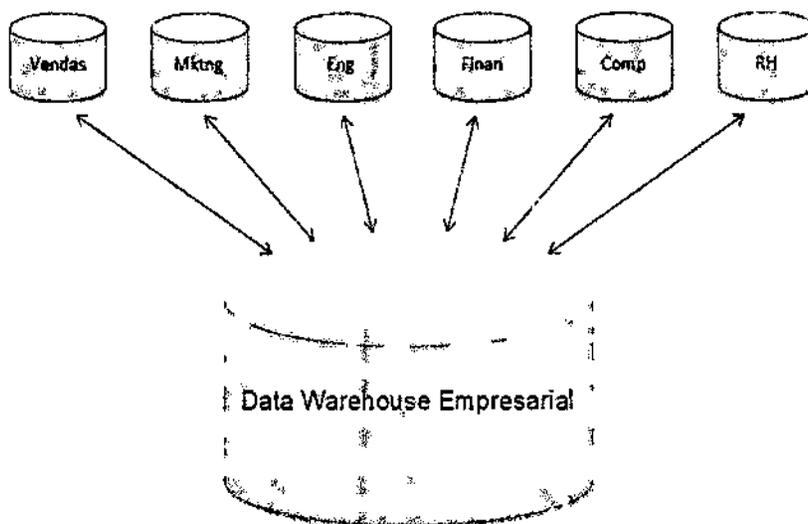


Figura 05: Data Marts Independentes em um Data Warehouse Empresarial; Fonte: Conrado Machado.

De acordo com Bain et. Al. (2000) conclui-se que:

Um Data Mart é uma estrutura de dados que se assemelha a de um DW e é projetado para ajudar com suporte à decisão, mas possui um subconjunto de dados de DW de uma forma que atenda aos requisitos específicos.

Uma das vantagens de um Data Mart é a sua mobilidade. Um vendedor, por exemplo, pode carregar o Data Mart em um laptop e levá-lo para uma de suas viagens de vendas. O Data Mart pode, neste caso, ser adaptado para oferecer informações que auxiliarão nas vendas em particular. Por exemplo, se a viagem é de vendas para o Japão, o Data Mart levado pode ser sobre as vendas no Japão durante os últimos anos, com informações sobre os clientes-chave e estatísticas sobre as operações que tenham guardado no Data Mart. Com isso, o vendedor pode fazer o seu trabalho e usar os dados no Data Mart para seu benefício, sem ter de estar ligado diretamente ao Data Warehouse da empresa.

2.7.1 Data Mart versus Data Warehouse

Existem duas diferenças principais entre um Data Mart e um Data Warehouse. A primeira é que o Data Mart contém um subconjunto dos dados no Data Warehouse. Os Data Marts são projetados para satisfazer uma necessidade específica, como dados do departamento, dados de um grupo de clientes, dados por um período determinado, etc. O Data Warehouse, por outro lado, serve como um repositório de dados da empresa, independentemente de requisitos específicos.

O Data Warehouse é projetado como parte da abordagem estratégica para o planejamento empresarial. Já o Data Mart é idealizado como parte do planejamento tático para satisfazer uma determinada exigência. Isso não significa, é claro, que uma empresa pode projetar os Data Marts de forma aleatória, sem olhar para a estratégia global. Se isso acontecer, os Data Marts resultantes serão desconectados e será muito difícil integrar o DW da empresa futuramente. Portanto, a concepção da estratégia de Data Warehousing deve incluir determinadas normas e critérios que os projetos de Data Mart deverão obedecer, de maneira que não comprometa a sua assimilação no DW posteriormente.

Conseqüentemente, o fato de que um Data Mart inclui apenas um subconjunto do Data Warehouse significa que a implementação do DW é mais

demorada e custa muito mais. A implementação do Data Mart é muito mais rápida e custa muito menos.

Muitos sistemas de suporte a decisão (DSS), vieram com "out-of-the-box" – soluções de Data Mart que afirmam poder ser implementadas num período muito curto de tempo – Depois de selecionar os requisitos específicos que não serão abrangidos por esses sistemas, é preciso muito esforço para personalizá-los a fim de que se construa o Data Mart. Além disso, com essas soluções, não é garantido que o design do Data Mart obedeça a tudo que foi definido nas diretrizes do projeto.

A outra diferença principal entre o Data Mart e o Data Warehouse é que os dados no DMRT podem ser mais detalhados do que no DW. Uma vez que os requisitos do DMRT são mais definidos do que os do DW, é possível pré-agregar os dados ao longo dos requisitos conhecidos e armazenar os resultados das consultas que se esperam ser executadas com o DMRT para assim fazer a extração dos dados de forma mais rápida e eficiente.

2.7.2 Tipos de Abordagens

Especialistas desenvolveram três principais abordagens para a construção de DMRT. Abordagens essas com sua peculiaridade e função específica que auxiliam o Data Warehouse nos processos de ETL.

Essas abordagens classificam-se em Top-Down, Bottom-Up e abordagem federada.

2.7.2.1 Abordagem Top-Down

Essa abordagem envolve a construção de um DW e sua povoação a partir de um conjunto de Data Marts dependentes de dados. Por exemplo, dados relacionados aos recursos humanos (RH) de departamento podem ser extraídos do DW em um dos Data Marts, e os dados pertencentes aos clientes podem ser extraídos em outro Data Mart. Os dados provêm de áreas locais para o DW, que, em seguida, o alimentam a partir dos Data Marts dependentes. Essa técnica de desenvolvimento tem como benefícios:

- Empregar uma metodologia rigorosa e familiar para o recolhimento, modelagem e implementação dos processos de requisição de decisão do

usuário final. Esta metodologia vem da implementação do DW, onde é passado para a compreensão e implementação desses DMRT.

- Criar um sistema de armazenamento de dados que oferece aos usuários finais a capacidade de ter uma perspectiva de toda as operações negociais da empresa, problemas e potenciais oportunidades de desenvolvimento de negócios.
- É baseado em um modelo de dados orientado por assunto que minimiza os problemas de integração entre os projetos de DW. Isso também é devido principalmente ao fato de que todos os Data Marts seguirão o mesmo design e implementação de processos, assim integrá-los mais adiante usando ferramentas DSS será uma tarefa relativamente fácil.
- Permite que Data Marts dependentes possam ser construídos a partir de informações da empresa, fornecendo assim uma abordagem adquirida usando tecnologia de Data Mart, conforme na figura 06.

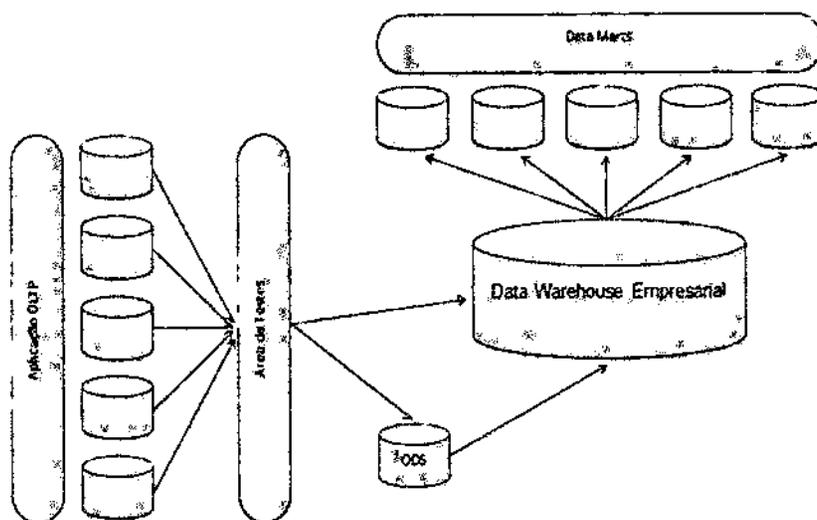


Figura 06: Abordagem Top-Down; Fonte: Conrado Machado.

No entanto, esses benefícios devem ser examinados em contrapartida às seguintes desvantagens antes de decidir seguir esta abordagem:

- A técnica de Top-down em Data Marts muitas vezes leva a prazos longos de entrega, alta capitalização, erros relacionados a custos e uma funcionalidade pobre no usuário final, mesmo quando as justificativas de custo adequado são feitas antes do projeto. Isso ocorre porque o projeto de Data Mart é dependente do Data Warehouse, que ainda pode estar evoluindo. Além disso, o orçamento do desenvolvimento de Data Mart é, na maioria dos casos,

dependente diretamente do Data Warehouse e de outros Data Marts ainda em desenvolvimento.

- Abordagens sobre Data Warehouses empresariais com os seus ciclos de longo alcance não podem oferecer soluções rápidas o suficiente para satisfazer a demanda das organizações nem soluções que lhes permitam responder rapidamente às mudanças nas condições de negócios e rapidamente explorar novas oportunidades de negócio.

2.7.2.2 Abordagem Bottom-Up

Esta abordagem envolve a construção de Data Marts em primeiro lugar, agregando, em seguida, seus dados a um DW. Na verdade, alguns dos defensores desta abordagem vão bem mais longe, afirmando que não há a necessidade de um Data Warehouse central na empresa. É mais pertinente ter uma espécie de Data Warehouse distribuída; este é composto de Data Marts independentes ligados através de software de integração de middleware ou ferramentas de apoio à decisão que tomam uma visão consistente de todos esses Data Marts. Estes Data Marts supostamente "colados" e independentes criam um ambiente de Data Warehouse distribuídos, e podem igualmente fornecer outros serviços importantes, como visões de negócio dos Data Warehouses, segurança reforçada, acompanhamento de instalações e melhor controle sobre o usuário final ao acesso dos dados.

Então, de acordo com essa abordagem, os fluxos de dados vindos diretamente do sistema operacional fluem diretamente aos Data Marts independentes. Estes podem, eventualmente, fornecer dados ao Data Warehouse como demonstrado na figura a seguir:

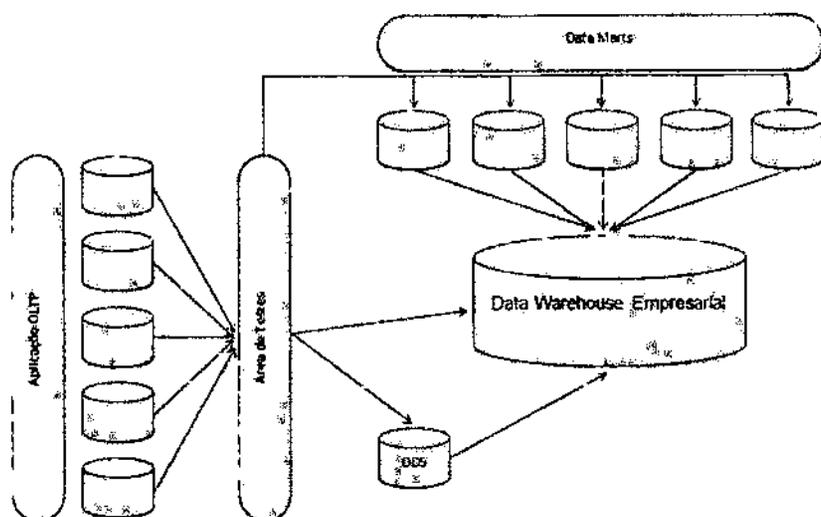


Figura 07: Abordagem Bottom-Up; Fonte: Conrado Machado.

Assim como a Top-Down, a abordagem Bottom-Up também possui suas desvantagens:

Uma construção e proliferação descontrolada em Data Marts independentes pode resultar em problemas de integração entre esses Data Marts e o Data Warehouse futuramente. A maioria desses problemas de integração deve-se, no entanto, às diferenças nos formatos de dados e representações entre os projetos de Data Marts. Estes problemas podem ser resolvidos através da utilização de um modelo comum de armazenamento de informações (um modelo de negócios ou um modelo de dados) para documentar os requisitos de processamento de decisão no desenvolvimento dos Data Marts.

Como os Data Marts se proliferam, os usuários corporativos desejarão acessar Data Marts pertencentes a outros departamentos da empresa. O acesso imediato aos Data Marts é difícil sem um middleware de banco de dados adequado, o que torna improvável garantir a transparência. É difícil controlar e pode até ocasionar um desempenho baixo e acesso com múltiplas consultas no Data Mart.

A necessidade de se desenvolver Data Marts independentes o mais rápido possível frequentemente leva ao entendimento de que Data Marts devem ser conduzidos pelos dados que existem em sistemas operacionais, ao invés de requisitos de negócios relacionados às informações de usuário. É importante ressaltar, porém, que essa crítica pode também ser nivelada em muitas implementações de Data Warehouses. Na verdade, em alguns projetos instalados de maneira rápida, um Data Mart de estrutura semelhante ao do sistema OLTP

fornecerá dados. Isto derrubará a finalidade de ter um Data Mart que segue os princípios de compreensão de um Banco de Dados dimensional como um Data Warehouse.

2.7.2.3 Abordagem Federada

Como já foi visto, as abordagens anteriores para implementação de Data Marts têm suas próprias forças e fraquezas. É preciso uma solução de meio termo para garantir o baixo custo e rápido retorno sobre o investimento (ROI) sem os problemas de integração de dados no futuro. Para isso, a concepção e desenvolvimento de Data Marts independentes devem ser geridas e baseadas em um modelo de informações comuns de requisitos de dados do Data Warehouse no processamento de decisão.

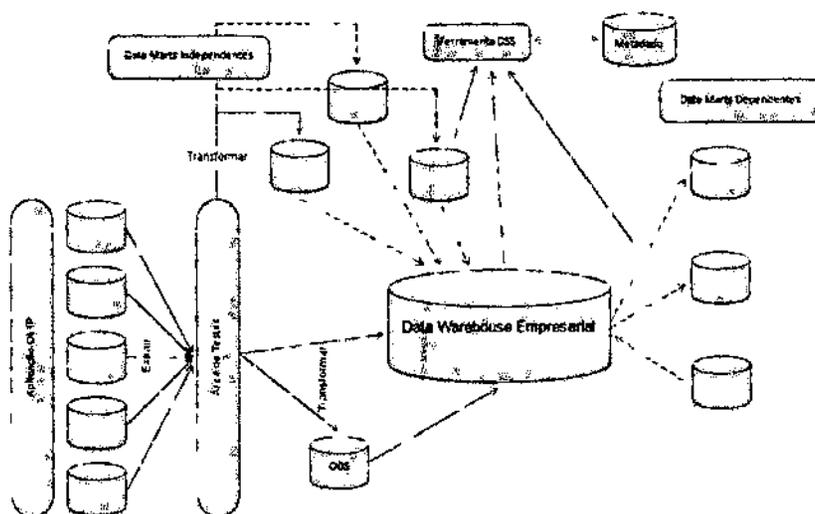


Figura 08: Abordagem Federada; Fonte: Conrado Machado.

Com a abordagem federada, o desenvolvimento do sistema de Data Warehousing torna-se um processo iterativo. O sistema de Data Warehouse no caso contém Data Marts independentes que começam os seus dados de feed diretamente na área de teste. O sistema também pode incluir um Data Warehouse empresarial tomando como base Data Marts dependentes que começam seu povoamento a partir do Data Warehouse da empresa, como se vê na figura acima.

A chave para a integração dos dados em um Data Warehouse Federado é um modelo comum de informações armazenadas e gerenciadas pelo sistema de armazenamento. Este modelo de informação é atualizado conforme novos Data

Marts independentes são construídos, ou quando novos Data Marts dependentes são adicionados a um DW.

Como se vê na figura, os dados nos sistemas operacionais são extraídos para serem carregados nos Data Marts independentes ou no Data Warehouse. Com os novos Data Marts independentes adicionados, ou novas áreas temáticas (Data Marts dependentes) adicionadas ao Data Warehouse, uma quantidade considerável de trabalho precisa ser feita para apoiar a extração e transformação para a nova adição. Com o modelo federado, essa preocupação tem sido reduzida pelo uso da área de teste, em que os dados operacionais são extraídos primeiro. Os dados são então transformados em uma forma que se adapte aos Data Marts ou Data Warehouses da empresa antes de serem carregados. Com o uso de um modelo de informação comum, pelo menos, a extração dos dados pode ser normalizada de forma que, a partir dessa perspectiva, faz novos requisitos e pequenas melhorias no processo de extração.

A partir da tabela acima, podemos gerar um gráfico demonstrador da quantidade de vendas de cada vendedor pela data e tirar as devidas conclusões:

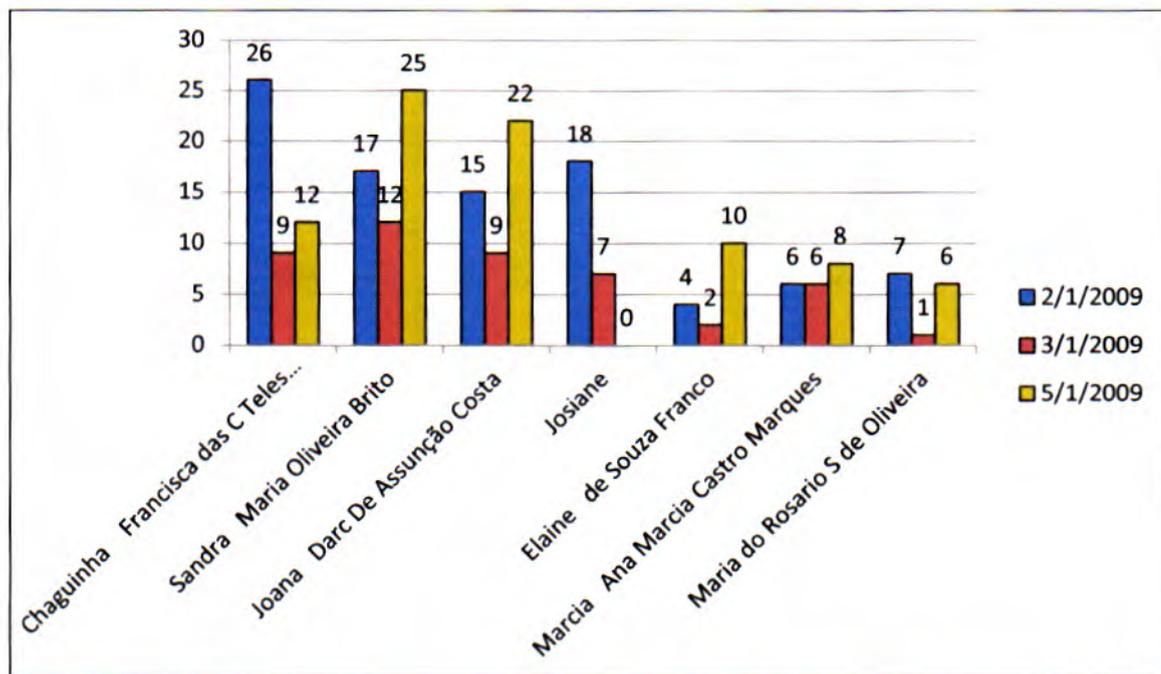


Gráfico 01: Quantidade de vendas realizadas no período de 02/01/2009 a 05/01/2009 por cada vendedor; Fonte: Dados obtidos na ferramenta de ETL.

Agora, para realizamos uma pesquisa mais completa que a anterior; extraímos o total de vendas de todas as funcionárias no ano de 2009, classificando o total de vendas por mês:

4 – CONSIDERAÇÕES FINAIS

Já sabemos que Data Warehouses são indispensáveis para que uma empresa considere-se competitiva e participativa no mercado global. Podemos afirmar isso por seu poder e contribuição que pode dar a uma empresa, seja ela qual for, pois possui um padrão de organização muito bem estruturado, além de ser apropriado para um alto conteúdo de buscas e consultas. A exploração de Data Warehouses se mostra de extrema importância ao suporte de Business Intelligence, independente de assunto ou área em questão. Pode-se dizer que Business Intelligence é o futuro para as organizações que desejam além de informação, conhecimento.

As novas tecnologias em ferramentas de extração e tomada de decisões estão para ajudar a melhorar não só o comércio pois estas não são voltadas apenas para essa atividade; em tudo que se usa a tecnologia de armazenamento de dados, pode-se aplicar técnicas de Data Warehousing, bem como as ferramentas que dependem de sua existência.

A intenção de trabalhar em cima do Data Warehouse utilizando a aplicação que realiza o processo de ETL para a empresa Magazine Rosemary Ltda foi de poder contribuir para a melhoria do comércio, organização de dados e assim poder melhorar o lucro.

Visando ampliar esse projeto como propostas futuras, o desenvolvimento do mesmo Data Warehouse por Data Marts e aplicação de Data Mining para assim melhorar mais ainda os negócios da empresa e migrar a ferramenta para uma linguagem livre, como Java, por exemplo.

Espera-se que este trabalho sirva de exemplo para apoiar o desenvolvimento de outros ambientes de Suporte à Decisão apoiados na tecnologia de Data Warehouse.

REFERÊNCIAS BIBLIOGRÁFICAS

- ALVES, R. P. L. **Data Warehouse Aplicado na Gestão de uma Telecom.** Uberlândia: UNIMINAS, 2006. 60 p.
- BAIN, Tony. Et. Al. **Professional SQL Server 2000 Data Warehousing with Analisys Services.** Wrox. Reino Unido. 2001.
- CASTRO, C. S. **Data Warehouse Estudo Comparativo das Ferramentas.** Uberlândia: UNIT, 2001. 69 p.
- ELMASRI, R.; NAVATHE, S. B. **Fundamentals of database systems.** California: Ed. Addison Wesley, 2000.
- FAYYAD, U. M. **Data mining and knowledge discovery: making sense out of data.** *IEEE Expert*, 1996.
- GOUVÊA, André Vinicius. Et. Al. **Uma aplicação de Data Warehouse para apoiar negócios.** Rio de Janeiro: UERJ. IME - Dept de Informática e Ciência da Computação.
- HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques.** Simon Fraser University. 2000.
- HUMPHRIES, Mark. **Data Warehousing: Architecture and Implementation.** [S.I.]. 1999.
- INMON, William Harvey. **Building the Data Warehouse.** 4. ed. Wiley. Indianapolis. 2005.
- KANASHIRO, Augusto. **Um Data Warehouse de Publicações Científicas: Indexação Automática da Dimensão Tópicos de Pesquisa dos Data Marts.** São Carlos: USP, 2007. 109 p.
- KIMBALL, Ralph. **Data Warehouse Toolkit.** São Paulo - Makron Books, 1998.
- SILBERSCHARTZ, Abraham; KORTH, Henry F.; SUDARSHA, S.. **Sistema de Banco de Dados.** 3. ed. São Paulo. Makron Books, 1999.
- LAROSE, Daniel T.. **Data Mining Methods and Models.** Wiley. New Jersey. 2006.
- MAHAJAN, Satish. **Building a Data Warehouse Using Oracle OLAP Tools.** ACTA Journal, 1997.
- RAINARDI, Vincent. **Building a Data Warehouse With Examples in SQL Server.** New York. Apress, 2008.
- SANTOS, I. M. **Data Warehouse Como Ferramenta de Auxílio em Sistemas de Monitoramento Ambiental.** Cuiabá: UFMT. Departamento de Ciência da Computação, 2003. 41 p.

SILVERS, Fon. **Building and Maintaing a Data Warehouse**. Florida.2008. Auerbach Publications.

VIEIRA, Carla de Cássia Nardelli. **Protótipo de um Sistema de Auxílio às Negociações Financeiras da FURB utilizando Data Warehouse**. Blumenau: Universidade Regional de Blumenau. 2001. 65 p.