

KDD E MINERAÇÃO DE DADOS:

Revisão em Data Warehouses

Prof. Ronaldo R. Goldschmidt

ronaldo@de9.ime.eb.br

rribeiro@univercidade.br

geocities.yahoo.com.br/ronaldo_goldschmidt

1

DATA WAREHOUSES – UMA VISÃO GERAL

- *Introdução*
- *Conceitos Básicos*
- *Modelagem Multidimensional*
- *Projeto de Data Warehouse*

2

DATA WAREHOUSES – UMA VISÃO GERAL

INTRODUÇÃO

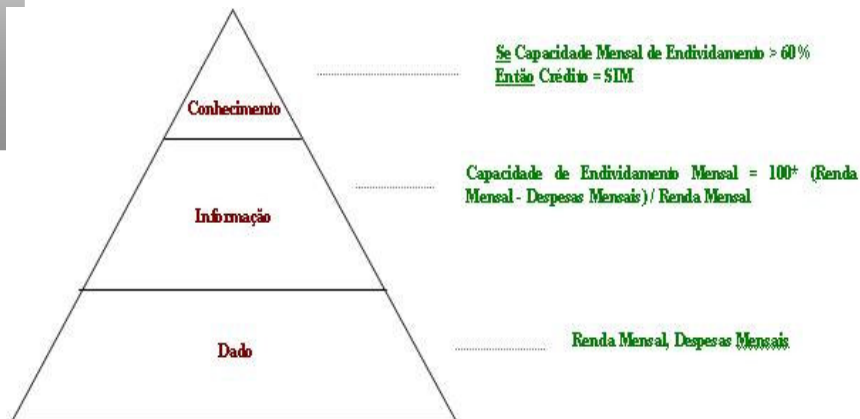
3

Avanços em TI e Crescimento BDs

- Internet
- SGBDs
- Dispositivos de memória (maior capacidade e menor custo)
- Leitoras de códigos de barras
- Sistemas de Informação em geral

4

Hierarquia: Dado - Informação - Conhecimento



5

Controles: Operacional e Estratégico

■ *Controle Operacional:*

- *Sistemas aplicativos transacionais de qualidade*
- *Automação de processos corporativos*

■ *Controle Estratégico:*

- *Busca de informações e conhecimento*
- *Planejamento corporativo*

6

Controle Operacional Corporativo

- *Ferramentas: OLTP*
- *OLTP – On-line Transactional Processing*
- *Modelagem de Dados para Sistemas OLTP*
- *Foco nos detalhes das transações do dia a dia*
- *Automação dos processos corporativos*

7

Controle Estratégico Corporativo

- *Ferramentas de BI – Business Intelligence (Inteligência de Negócios)*
- *BI – Conjunto de tecnologias orientadas a disponibilizar informação e conhecimento corporativos: CRM, KM, DW*
- *Foco em aspectos estratégicos para os negócios corporativos*
- *Requer bases de dados históricas*

8

DW - Data Warehouse

- *Armazém de Dados*
- *Principal objetivo: disponibilizar informações para apoio a decisões em empresas.*
- *Deve disponibilizar dados sobre a história da empresa, viabilizando consultas e análises estratégicas.*
- *Consultas OLAP – On-line Analytical Processing: permitem visualização e navegação pelos dados sob diversas perspectivas e níveis de detalhe.*

9

Aplicações Típicas

- *Pesquisa de fraudes;*
- *Análise de crédito;*
- *Análise de sazonalidade;*
- *Análise de risco;*
- *Integração de Informações de Clientes;*
- *Rentabilidade de Clientes e Produtos;*
- *Análises de Resultados de Vendas;*
- *Análises de Ações de Marketing.*

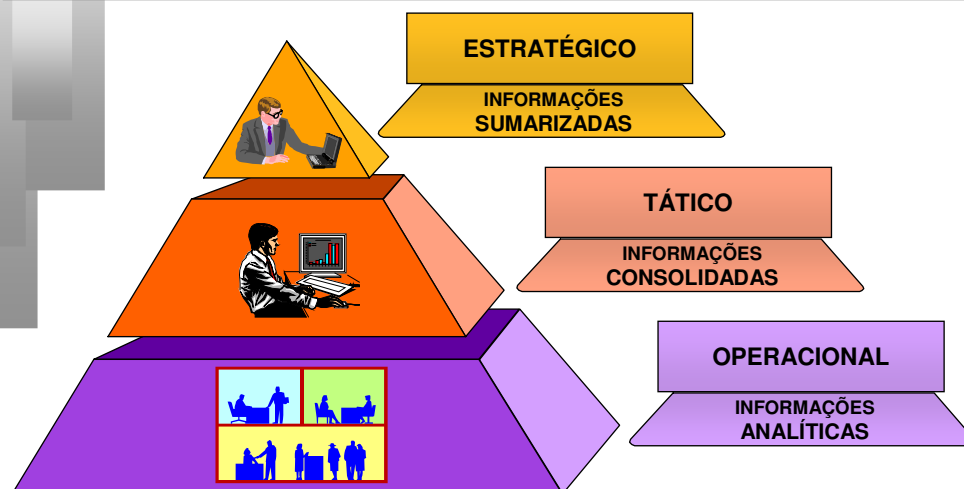
10

DATA WAREHOUSES – UMA VISÃO GERAL

CONCEITOS BÁSICOS

11

Perfil do Usuário X Tipo de Informação



12

Sistemas OLTP X OLAP

	OLTP	OLAP
Objetivo	Controle dos Proc. Operacionais	Tomada de Decisão
Cliente	Pessoal Operacional	Gestores do Negócio
Dados	Atômicos, Atualizados e Dinâmicos	Atômicos, Consolidados, Históricos e Estáveis
Estrutura	Normalizada	Dimensional
Tempo de Resposta	Segundos	De segundos a minutos
Foco	Orientado à Aplicação	Orientado à Informação
Acesso	Alto	Moderado a baixo
Atualização	Contínua	Periódica
Aplicação	Estruturada e processamento repetitivo	Não estruturada e processamento analítico

13

Data Warehouse

■ *Data Warehouse é uma coleção de dados*

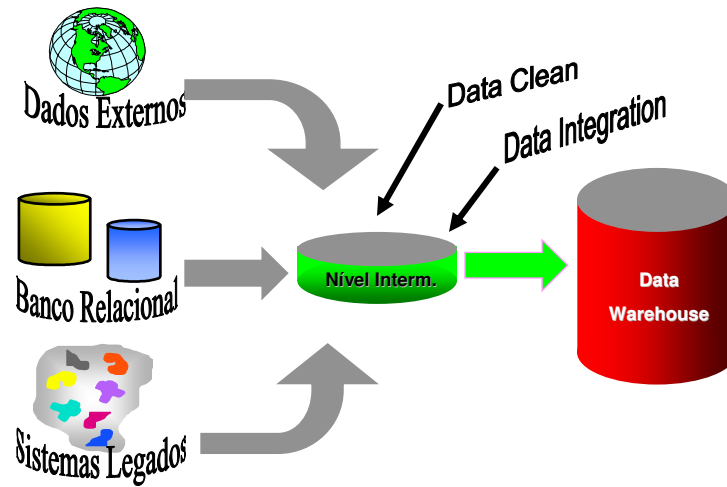
- *orientados a assuntos,*
- *integrados,*
- *não voláteis,*
- *variáveis com o tempo,*

destinados a auxiliar decisões de Negócio.

Willian H. Inmon - 1990

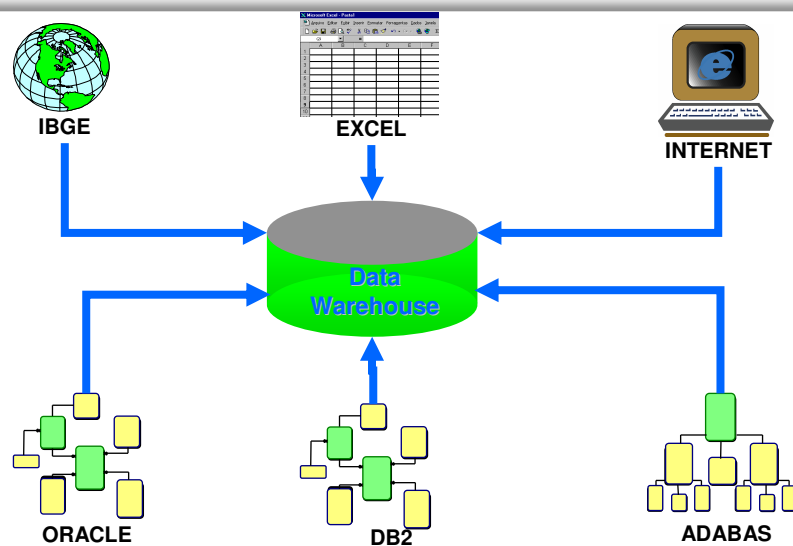
14

Exemplo de Ambiente



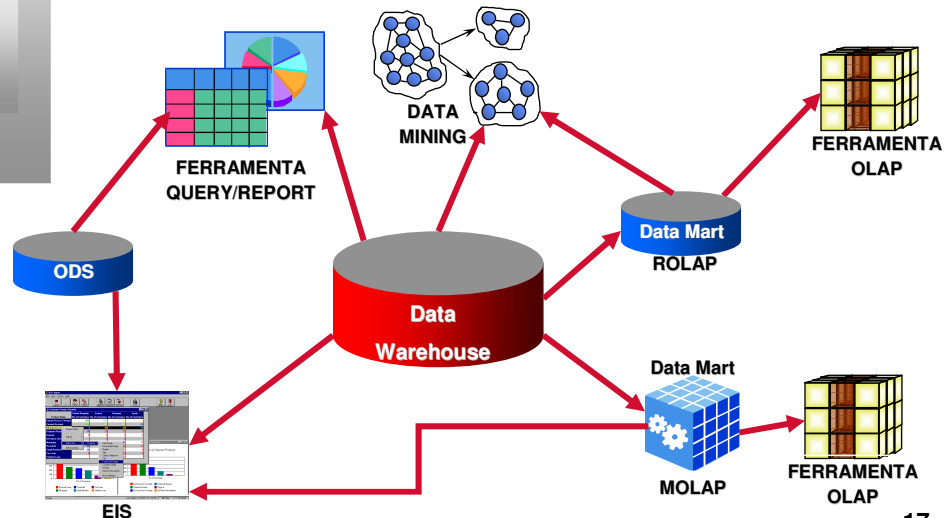
15

Exemplos de Fontes de Dados

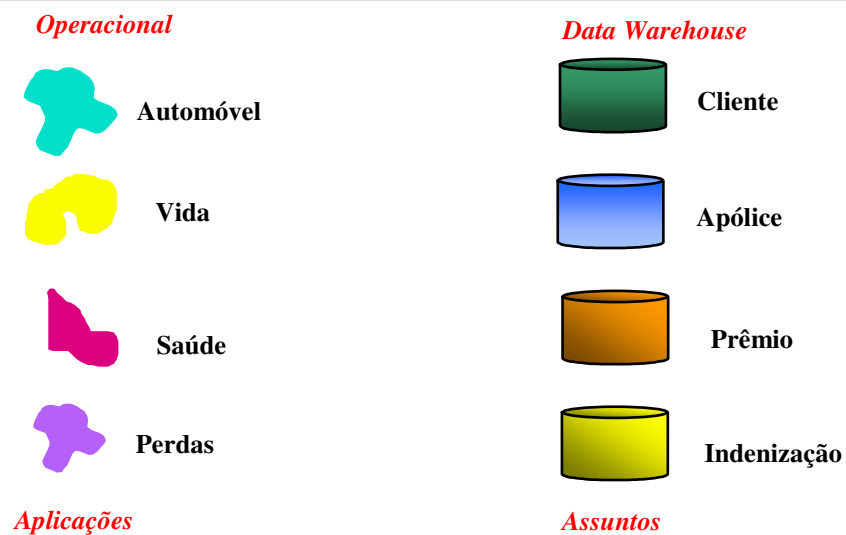


16

Exemplos de Usos do Data Warehouse



Característica de Orientação a Assuntos



Característica de Integração dos Dados



*Mesmos dados,
nomes diferentes*



*Dados diferentes,
mesmo nome*



*Dados só
encontrados aqui*

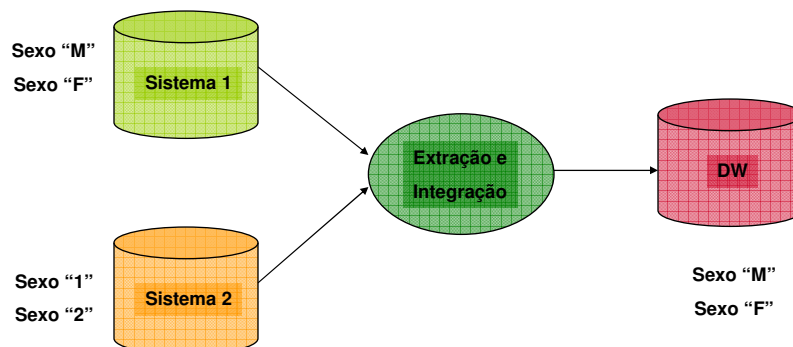


*Chaves diferentes,
mesmo dados*

- *A passagem de dados do ambiente operacional (legado) para o DW não é tão simples quanto uma mera extração.*

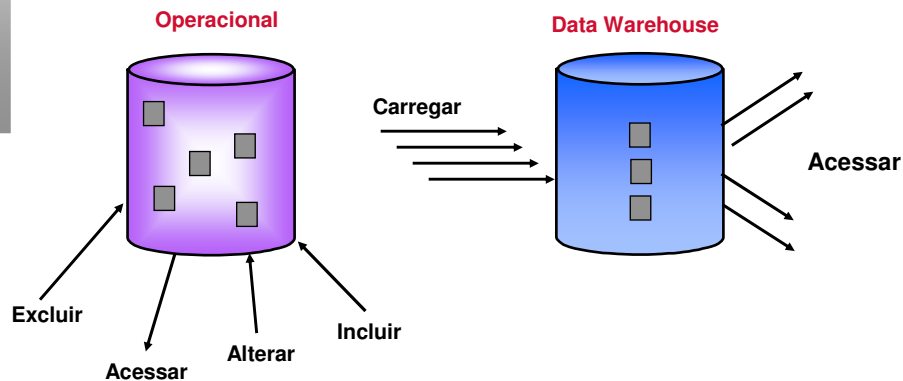
19

Característica de Integração dos Dados



20

Característica de Não Volatilidade



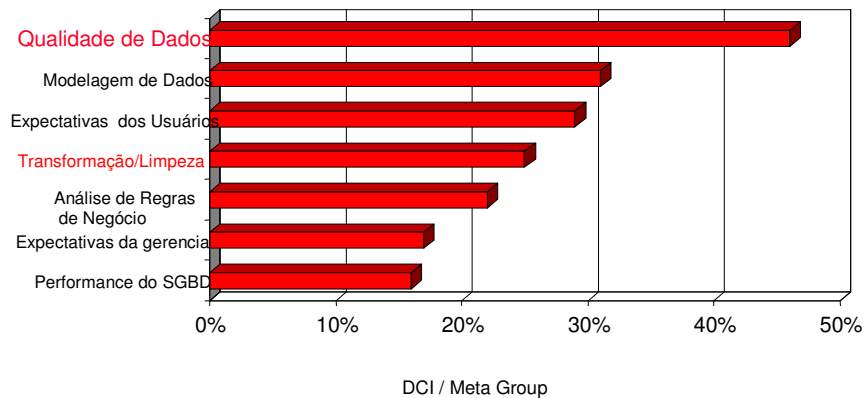
21

Fatores Críticos de Sucesso para DWs

- *Patrocinadores;*
- *Metodologia;*
- *Boa seleção de dados;*
- *Abordagem e ambiente adequados;*
- *Plataforma de hardware e software;*
- *Treinamento da Equipe de IT;*
- *Ferramentas de administração.*

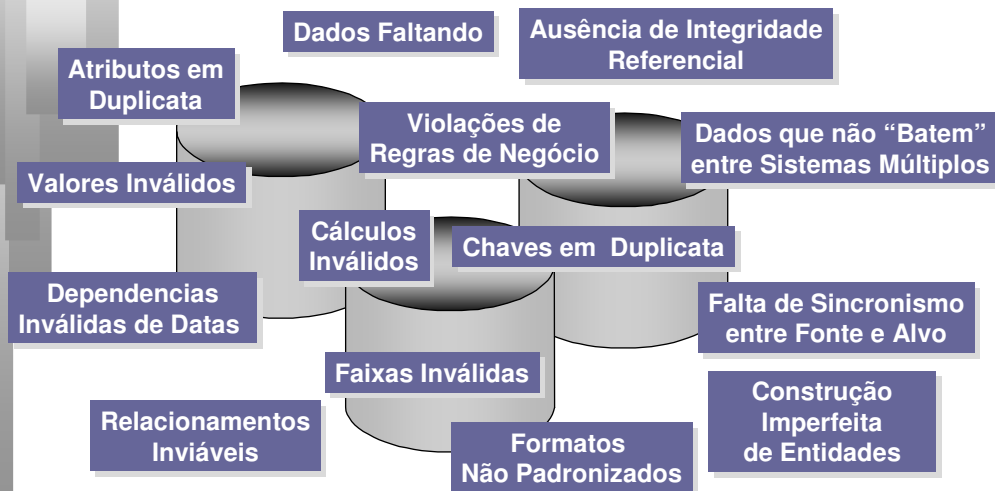
22

Maiores Desafios em Data Warehousing



23

Qualidade de Dados



24

Sumarização dos Dados

Alto nível de detalhes
Baixo nível de granularidade
(grãos menores)

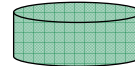


Exemplo:
Detalhe de cada venda
realizada por um vendedor
durante um mês

Data
Hora
Vendedor
Valor

100 registros por mês

Baixo nível de detalhes
Alto nível de granularidade
(grãos maiores)



Exemplo:
Sumário das vendas
realizadas por um vendedor
durante um mês

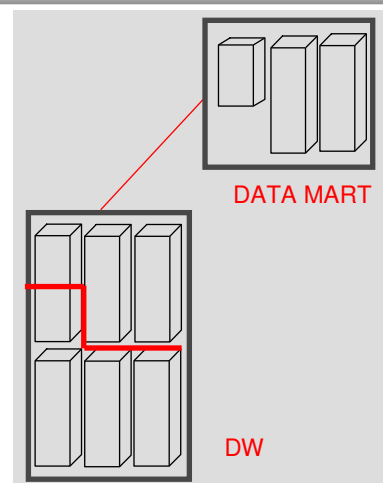
Mês
Vendedor
Valor

1 registro por mês

25

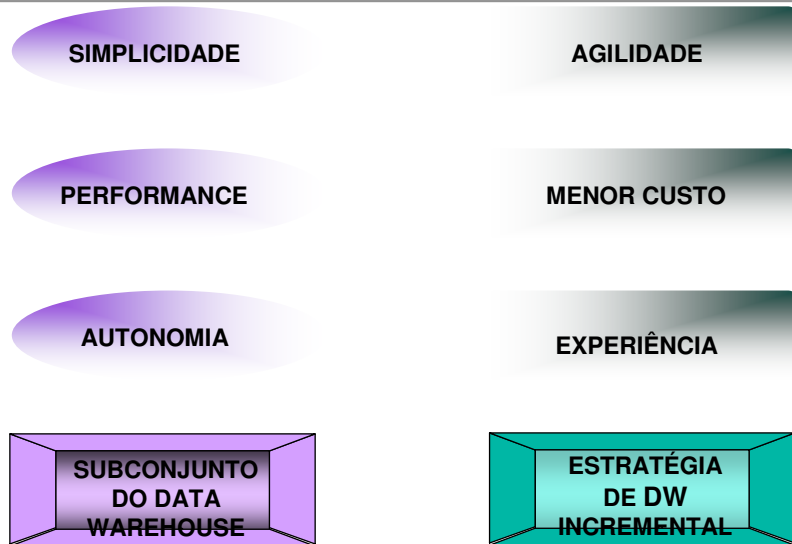
Data Mart

- *Data Mart é uma porção física ou lógica do Data Warehouse para atender a uma área da empresa.*



26

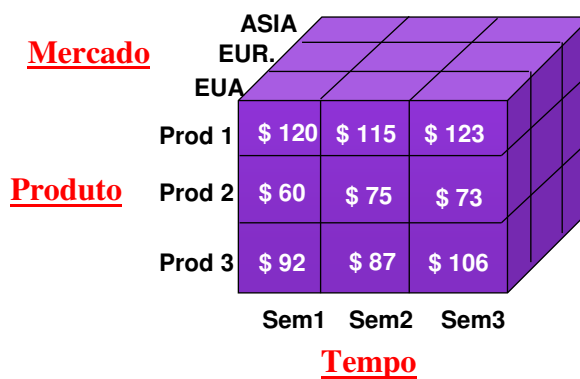
Razões para criação de um Data Mart



27

Cubo/Hipercubo de Dados

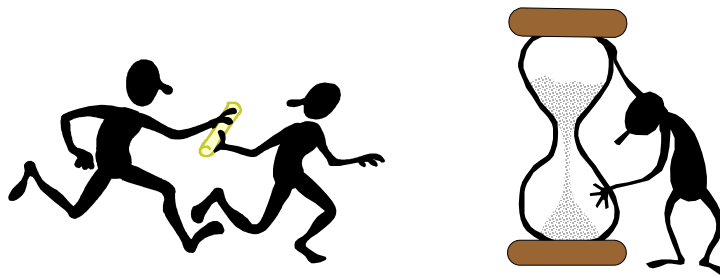
Recurso para cruzamento/visualização dos dados em aplicações OLAP.



28

Ferramentas de ETL/ETC

São as ferramentas do mercado que facilitam o processo de **Extração, **T**ransformação e **C**arga dos dados, bem como a análise e garantia da qualidade dos dados.**



29

Fontes de dados/informações

- *Em qual plataforma esta armazenada cada fonte de dados ?*
- *Qual sua periodicidade de atualização ?*
- *Tenho acesso a todas as fontes em um único ambiente ?*
- *Como as bases externas chegam à empresa e em que meios físicos ?*
- *Criação interfaces para informações gerenciais que não estão armazenadas nas bases de dados operacionais .*

30

DATA WAREHOUSES – UMA VISÃO GERAL

MODELAGEM MULTIDIMENSIONAL

31

MODELAGEM MULTIDIMENSIONAL

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

32

MODELAGEM DE DADOS

■ *Tradicional:*

- *Assegura Restrições de Integridade*
- *Evita redundância de informação*

■ *Multidimensional*

- *Facilita a realização de consultas “ad-hoc” por usuários não técnicos*
- *Acelera o desempenho de consultas*
- *Admite redundância de informação*

33

Visão Multidimensional – Um Exemplo



34

Visão Multidimensional – Um Exemplo

Papel	Bolsa	Mês	Lucratividade
TEL PN	Rio de Janeiro	Jan	+5%
TEL PN	Rio de Janeiro	Feb	-2%
TEL PN	Rio de Janeiro	Mar	+7%
TEL PN	São Paulo	Jan	+4%
TEL PN	São Paulo	Fev	-1%
TEL PN	São Paulo	Mar	+4%
PET PN	São Paulo	Jan	+2,5%
BB PN	Rio de Janeiro	Jan	-1%

35

MODELAGEM MULTIDIMENSIONAL

■ *Um modelo multidimensional possui três elementos básicos:*

- *Fatos*
- *Dimensões*
- *Medidas (Variáveis)*

36

Fatos - Definição

Fato

Coleção de itens de dados, composta de dados de medida e de contexto.

Representa um item/transação/evento de negócio.

Ex.

(Papel, Bolsa, Mês, Lucratividade)

(Tel PN, RJ, Jan, 2004)

37

Fatos - Características

- *Medidas de negócio, indicadores.*
- *Dados necessários para análise do negócio.*
- *Representados por valores numéricos.*
- *Evolutivos no tempo.*

38

Dimensões - Definição

Dimensão

Elemento que participa da definição de um fato.

Ex.

Papel, Bolsa, Mês, etc...



39

Dimensões - Características

- *Determinam o contexto do assunto.*
- *Normalmente não são representadas por valores numéricos.*
- *São descritivas/classificadoras.*
- *Pode conter membros hierarquicamente organizados.*

40

Identificação de Dimensões

Quando ?

1996						
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Onde ?



Quem ?



O que ?



Estas perguntas ajudam-nos a definir as dimensões

41

Membros de uma Dimensão

São utilizados para classificar dados dentro de uma dimensão.

Ex: Dimensão Geográfica (Onde?)

REGIÃO

UF

CIDADE

(Ano, Produto, Região, UF, Cidade, Qtde)

42

A Dimensão Tempo (Quando ?)

- *Sempre presente - DW é série temporal*
- *Deve ser explícita para separar dados:*
 - entre dias úteis e feriados*
 - em fins de semana*
 - em períodos fiscais*
 - em estações*
 - em eventos importantes (copa do mundo, falecimento Sena, Olimpíadas, etc)*

43

Medidas - Definição

Medida

Atributo ou variável numérica que representa um fato.

Ex.

Valor das Vendas, Número de Evasões, Quantidade de Produtos, etc...

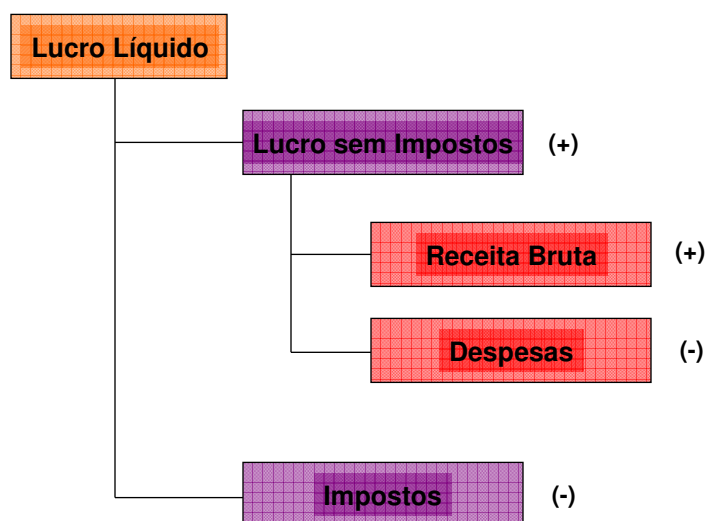
44

Medidas - Características

- *Indicadores de Negócios.*
- *São representadas por valores numéricos.*
- *Determinada pela combinação das dimensões que participam de um fato.*

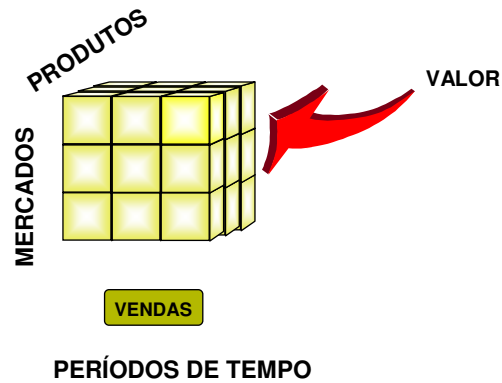
45

Hierarquia de Medidas - Exemplo



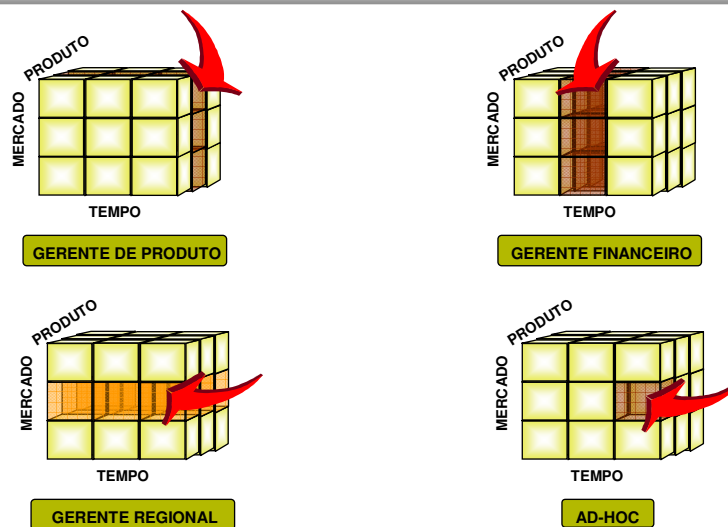
46

Modelo Multidimensional - Visualização



47

Modelo Multidimensional - Visualização



48

Modelo Multidimensional - Vantagens

- Modelo mais natural para o usuário, o que leva a uma facilidade de “navegação” pelas informações
- Acesso a um fato diretamente
- O mesmo conjunto de informações associadas a um fato pode ser visto sob várias óticas (dimensões), de forma simples e ágil

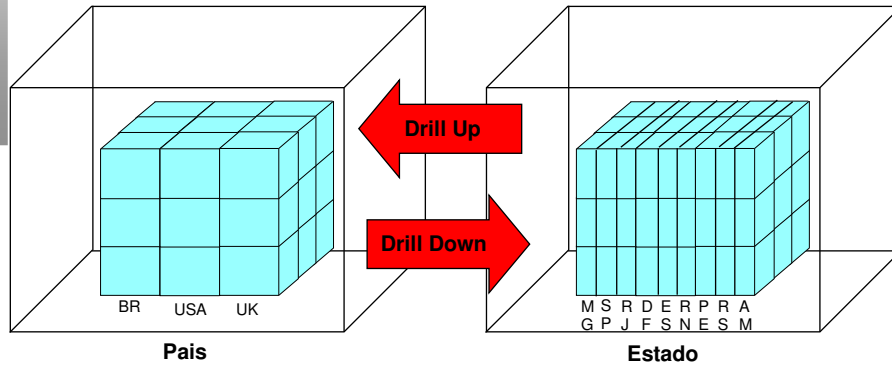
49

Operações Básicas em ferramentas OLAP

- Capacidade para fazer automaticamente:
 - Drill
 - Slicing
 - Dicing
 - Pivoting
 - Data Surfing
 - Consultas Ad-Hoc

50

Drill

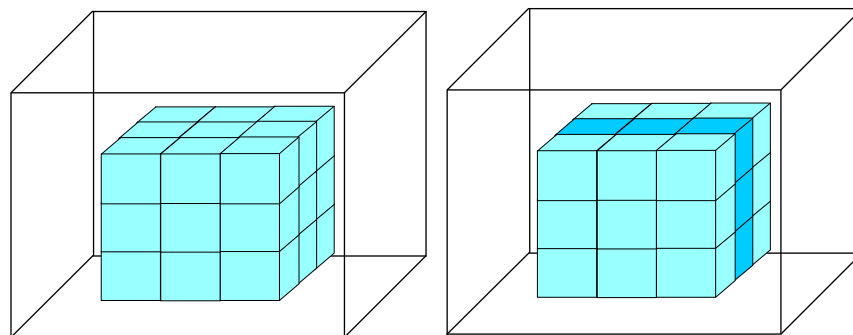


Aumentar ou Diminuir o Nível de Detalhe

Ex: Vendas por País ↔ Vendas por Estado

51

Slicing

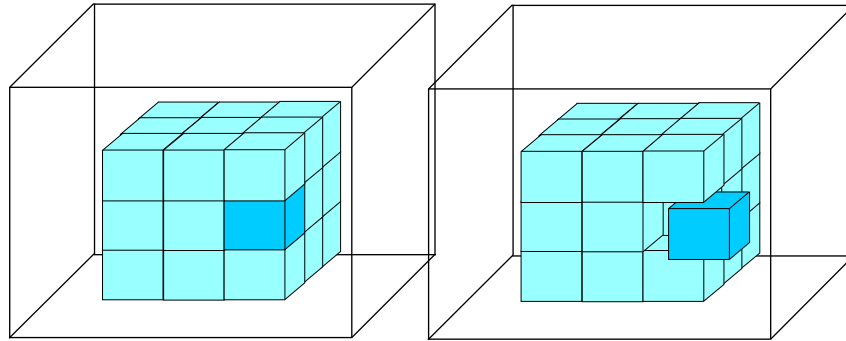


Selecionar as Dimensões para Consulta

Ex: Vendas por País por Mês (Somatório dos Produtos)

52

Dicing

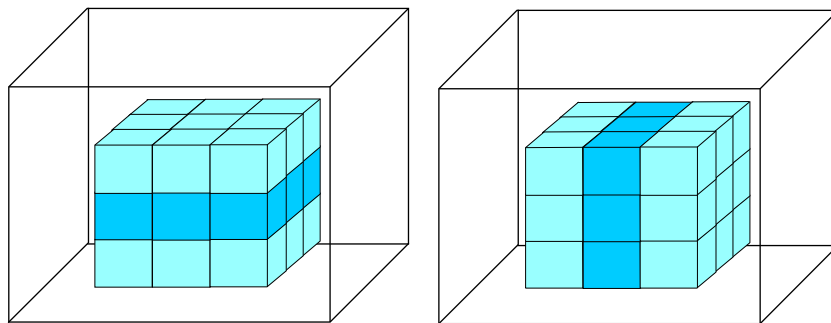


Limitar Conjunto de Valores algumas Dimensões

Ex: Vendas no Estado de Minas (por Produto por Ano)

53

Pivoting

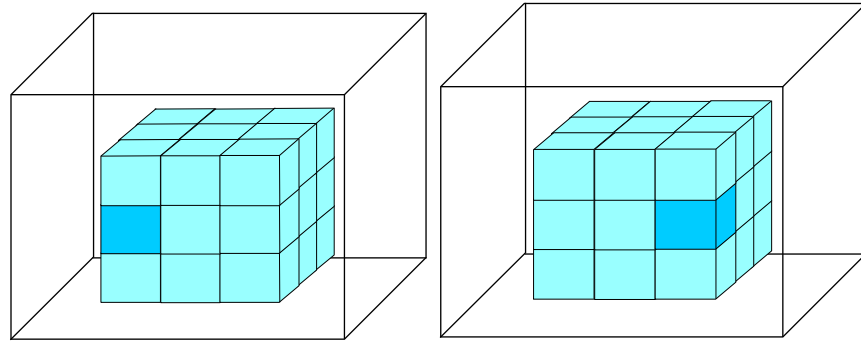


Trocar as Dimensões entre Linhas e Colunas

Ex: Vendas por Produto por Estado $\longleftrightarrow$ por Estado por Produto

54

Data Surfing



Executar a mesma análise em outro Conjunto de Dados

Ex: Vendas no Brasil $\longleftrightarrow$ Vendas no Reino Unido

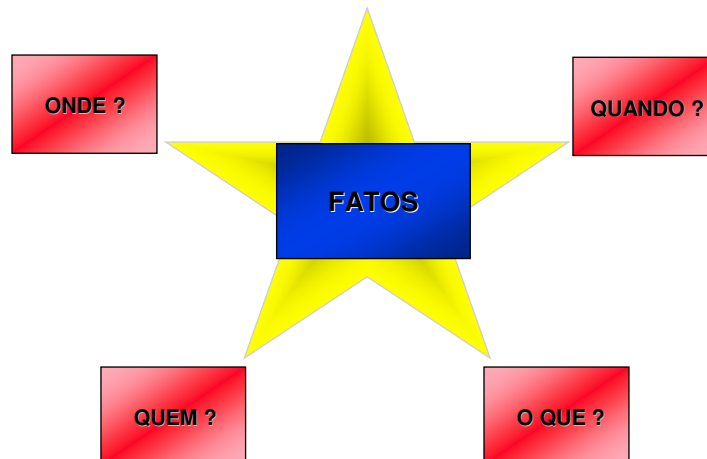
55

Consultas Ad-Hoc

- *Qualquer consulta que não seja pré-definida (Consulta Eventual)*
- *São consultas que unem os recursos oferecidos pela ferramenta utilizada às necessidades eventuais do usuário.*

56

Star Schema - Modelo Estrela



57

Modelo Estrela - Exemplo



Dimensão Cliente
Código Cliente
Nome Cliente
Atividade
Cidade
Estado
País
Segmento

Fato Vendas
Data
Código Vendedor
Código Produto
Código Cliente
Valor da Venda
Quantidade
Margem
Margem %

Dimensão Vendedor
Código Vendedor
Nome Vendedor
Código Região
Nome Região

**Uma tabela
para cada
dimensão**

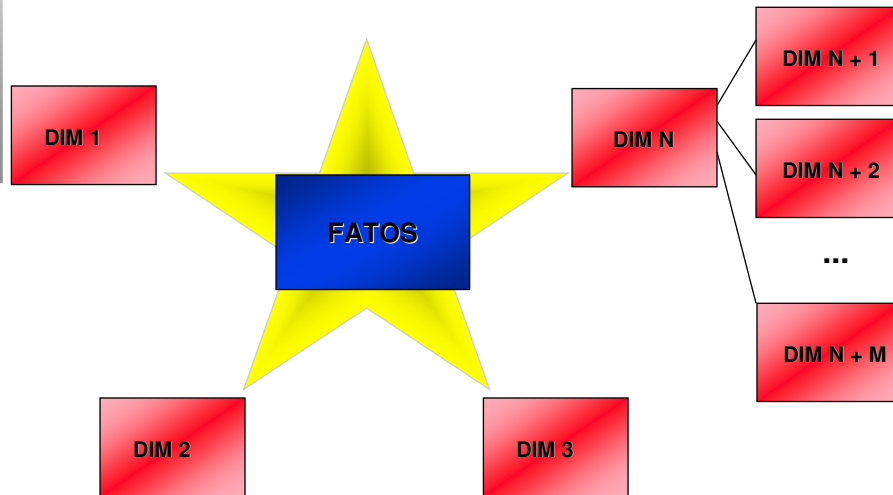
Dimensão Produto
Código Produto
Nome Produto
Código Grupo
Nome Grupo

Vantagens

- Reduz o número de joins nas consultas
- Simplifica a visão do modelo de dados
- Aumenta a performance das consultas

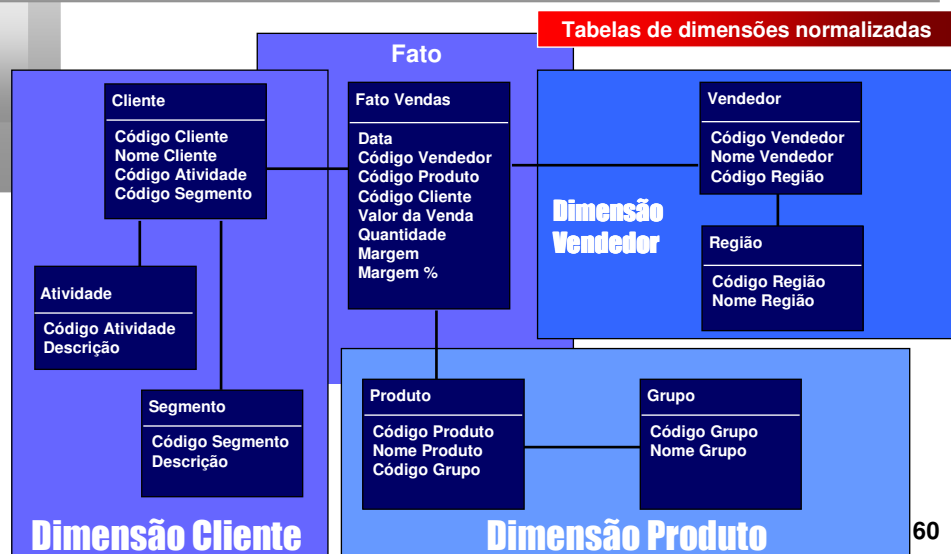
58

Snowflake Schema - Modelo de Flocos de Neve



59

Snow Flake Schema



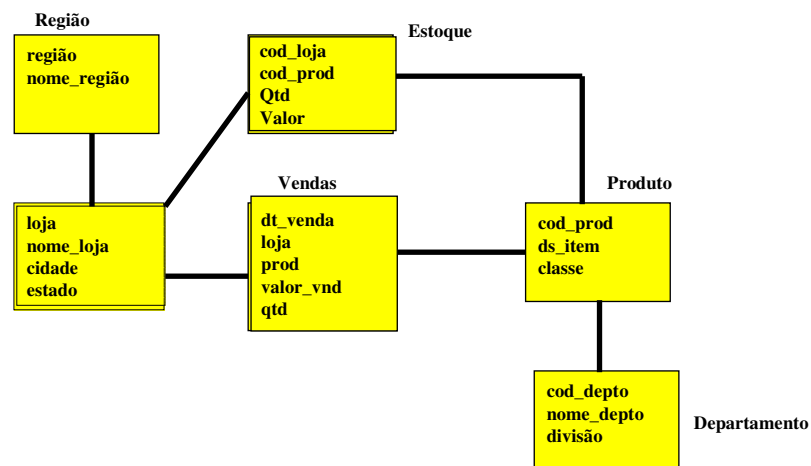
60

Ex. de Aplicação: Vendas no Varejo

- *Modelo de Dados OLTP*
- *Modelo de Dados Multidimensional*

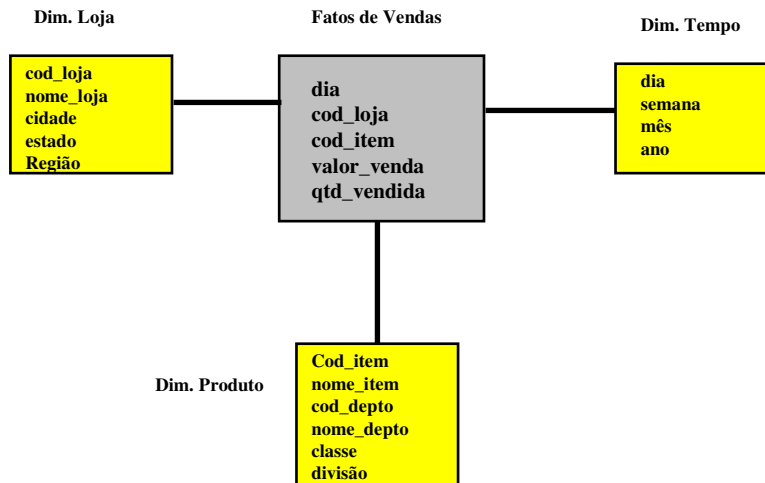
61

Ex.: Modelo de Dados OLTP



62

Ex.: Modelo Multidimensional



63

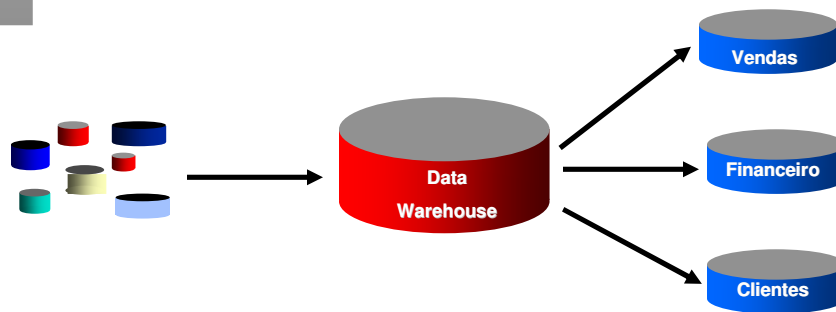
DATA WAREHOUSES – UMA VISÃO GERAL

PROJETO DE DATA WAREHOUSES

64

Abordagem Top Down

- *Bom controle*
- *Péssimos prazos*



65

Abordagem Top Down

Modelo extremamente ambicioso visando atender todas as necessidades corporativas.

Vantagens:

- ***Homogeneização das informações;***
- ***Capacidade para grande volume de informações;***
- ***Minimiza enormemente a redundância de dados;***
- ***Informações gerenciais em um único ambiente físico;***
- ***Gerenciamento centralizado da informação.***
- ***Processo mais longo: os requisitos podem se modificar até o momento da implementação.***

66

Abordagem Top Down

Modelo extremamente ambicioso visando atender todas as necessidades corporativas.

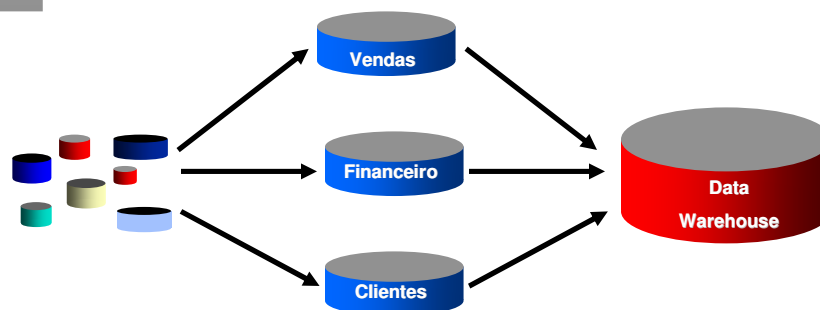
Desvantagens:

- **Maior tempo para a obtenção de resultados (~ 2 anos);**
- **Dificuldade de administrar : envolvimento de um grande número de pessoas, principalmente executivos;**
- **Difícil alcançar consenso em questões conceituais;**
- **Grande investimento inicial;**
- **Processo mais longo: os requisitos podem se modificar até o momento da implementação.**

67

Abordagem Bottom Up

- *Delivery rápido*
- *Ilhas de informações dificilmente integráveis*



68

Abordagem Bottom Up

Abordagem “Data Mart”

Vantagens:

- **Menor tempo para obtenção de resultados (3 meses);**
- **Esforço mais fácil de gerenciar;**
- **Bom ponto de partida para DSS's mais sofisticados;**
- **Aumento de performance a partir da experiência;**
- **Custo mais baixo e menos arriscado no curto prazo.**

69

Abordagem Bottom Up

Abordagem “Data Mart”

Desvantagens:

- **Duplicidade de dados e processos de extração;**
- **Perda do controle por parte do órgão central de TI;**
- **Expansibilidade e capacidade de armazenamento limitados;**
- **Não resolve diferenças conceituais entre áreas;**
- **Limita a capacidade de cruzamento de informações entre áreas, podendo gerar novas “ilhas” de informação;**
- **Integrar depois é difícil (de 3 a 4 vezes mais caro);**

70

Dez Grandes Erros

- Começar com patrocinadores corporativos errados;
- Criar expectativas que não podem ser realizadas;
- Carregar o warehouse com informações “simplesmente por estarem disponíveis”;
- Acreditar que um projeto de DW é igual a um projeto OLTP;
- Escolher um gerente do DW que é voltado à tecnologia ao invés de ser voltado ao usuário;

71

Dez Grandes Erros

- Concentrar-se nos dados tradicionais, ignorando dados externos, textuais, imagens, sons e vídeos;
- Disponibilizar dados com definições confusas e sobrepostas;
- Acreditar em promessas de performance, capacidade e escalabilidade;
- Acreditar que quando o DW estiver em operação, todos os problemas estarão terminados;
- Acreditar que só existem 10 grandes erros.

72

Bibliografia

- **Como usar o Data Warehouse**
 - W. H. Inmon e Richard D. Hackathorn
- **Projeto de Data Warehouse**
 - Felipe Neri
- **Como Construir o Data Warehouse**
 - W. H. Inmon
- **Data Warehouse Toolkit**
 - Ralph Kimball
- **Intranet Data Warehouse**
 - Rick Tanler

73

Referências na Internet

- **<http://www.idwa.org>** - International Data Warehousing Association;
- **<http://www.dw-institute.com>** - The Data Warehouse Institute;
- **<http://www.olapcouncil.org>** - OLAP council;
- **<http://www.rkimball.com>** - Ralph Kimball;
- **<http://www.tpc.org>** - Transaction Processing Council;
- **<http://www.olapreport.com>** - The OLAP Report.

74