

TSW017

Grandes Sistemas de dados - projeto e gerenciamento

Prof. Dr. Jorge Rady

1

Data Warehouse

Modelagem Conceitos básicos

2



Necessidade de
Disponibilidade



Dados são
distribuídos



Desejável bom
Desempenho



Processamento
transacional



Importância da
Segurança



Prazo para
Armazenamento
dos Dados

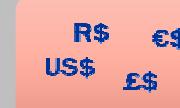


Exigida
Credibilidade



R\$ €\$
US\$ £\$
Dados não Padronizados

Motivação



Idéia do
**Data
Warehouse**

3

4

Data Warehouse

Data Warehouse é uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo e não voláteis, para dar suporte ao processo gerencial de tomada de decisões. (Inmon)

Data Warehouse é um processo em andamento que aglutina dados de fontes heterogêneas, incluindo dados históricos e dados externos para atender à necessidade de consultas estruturadas e ad-hoc, relatórios analíticos e de suporte à decisão. (Harjinder)

5

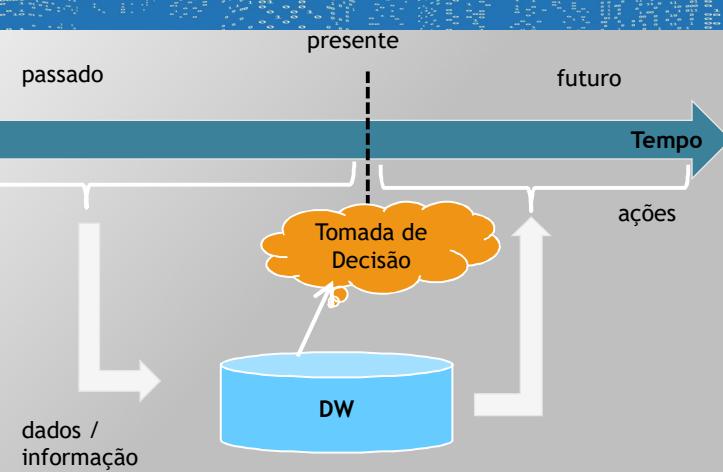
Data Warehouse

Data Warehouse é o processo de integração dos dados corporativos de uma empresa em um único repositório, a partir do qual os usuários podem facilmente executar consultas, gerar relatórios e fazer análises. (Singh)

Data Warehouse é o um sistema que extrai, limpa, trata e entrega dados de várias fontes em um modelo dimensional e suporta/implementa consultas e análises para o processo de tomada de decisão. (Kimball)

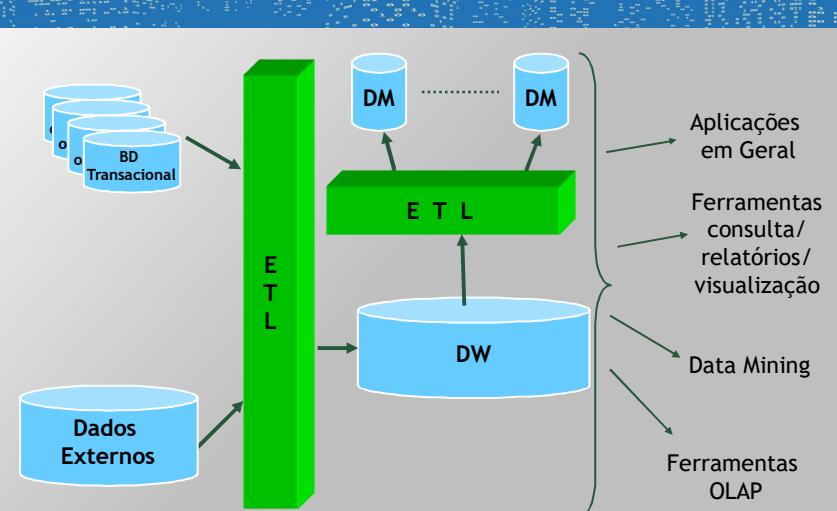
6

Data Warehouse

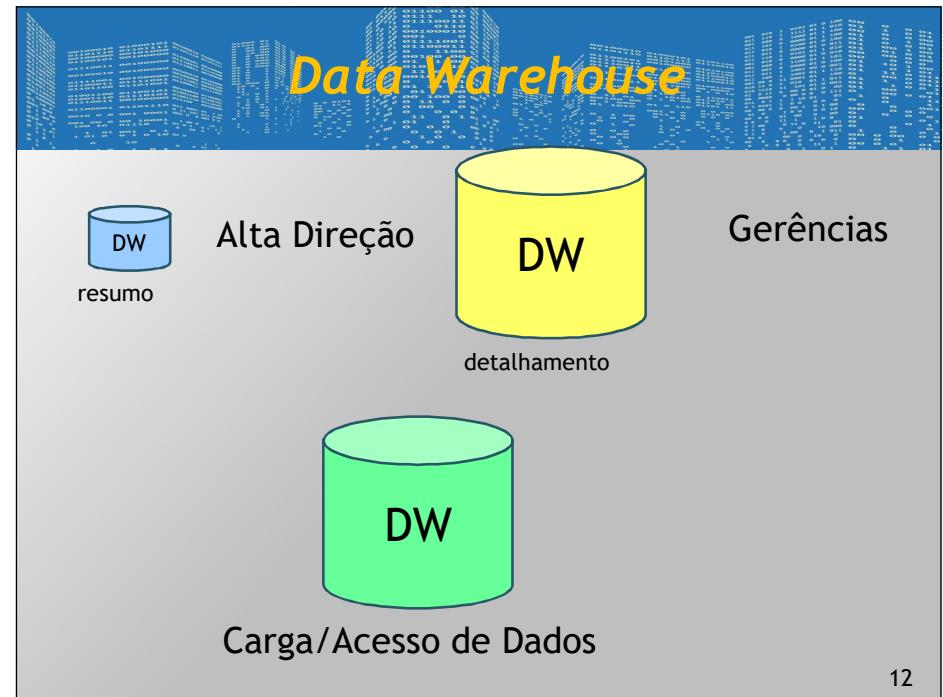
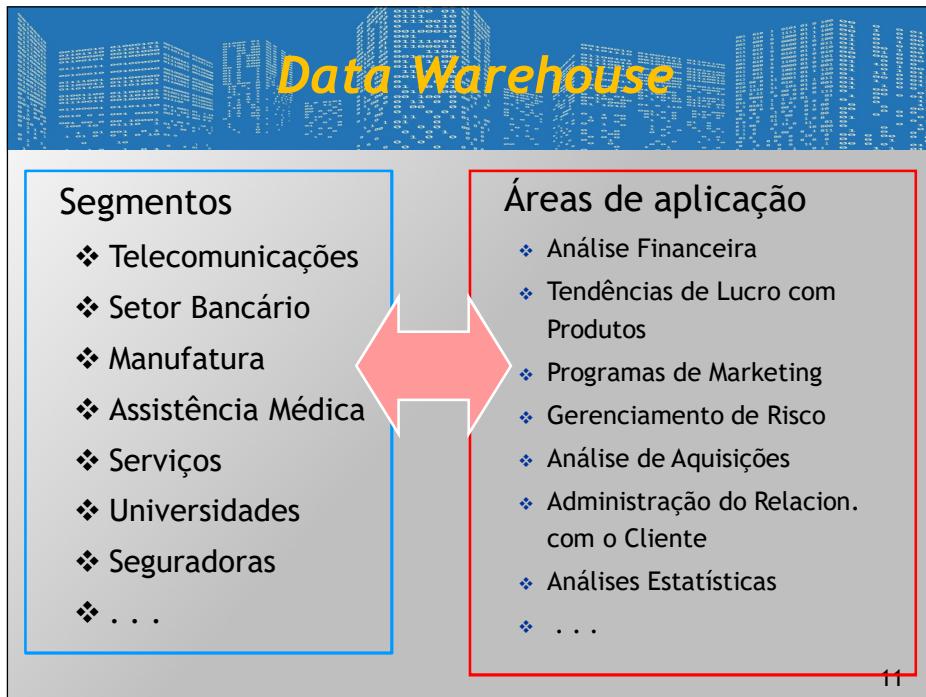
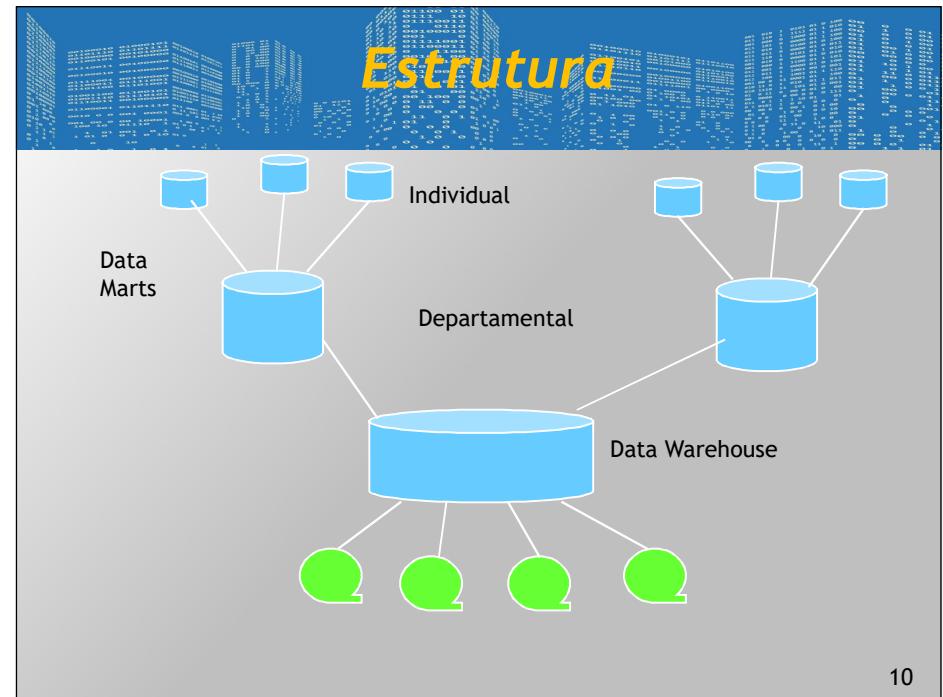
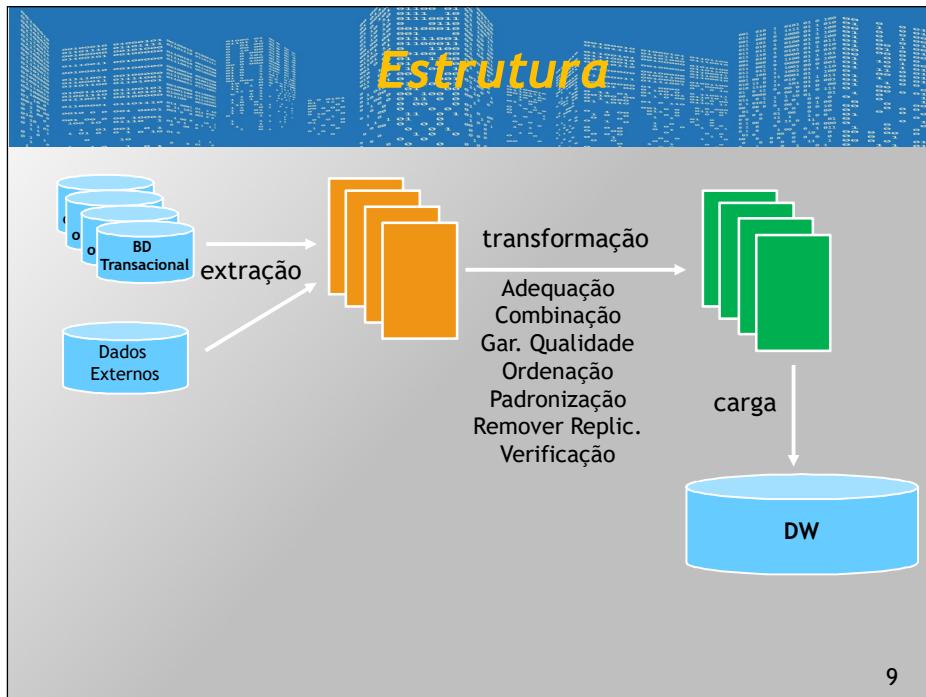


7

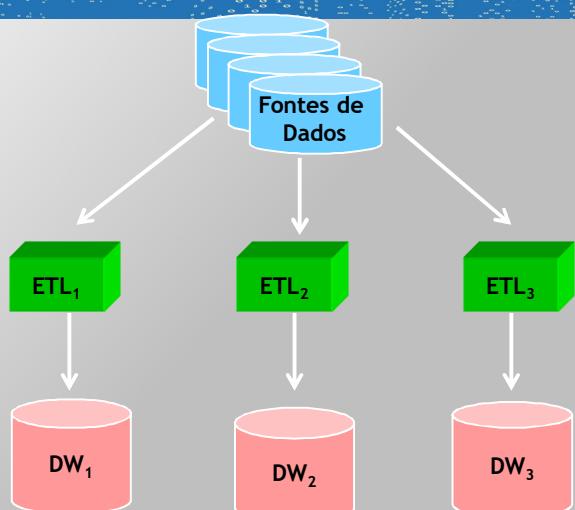
Estrutura



8

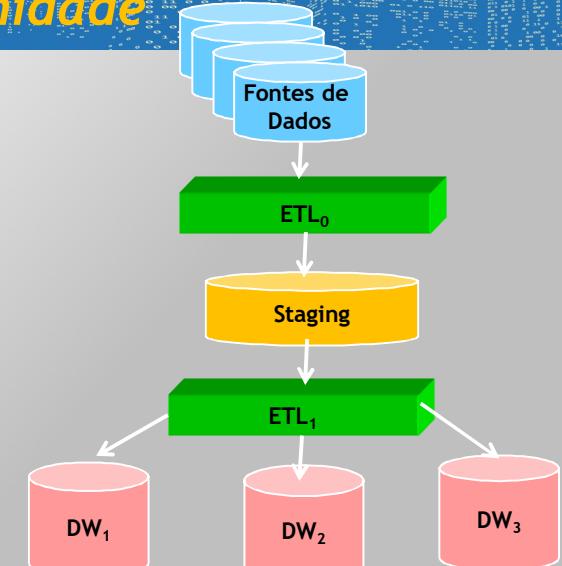


Arquiteturas - Data Marts Independentes



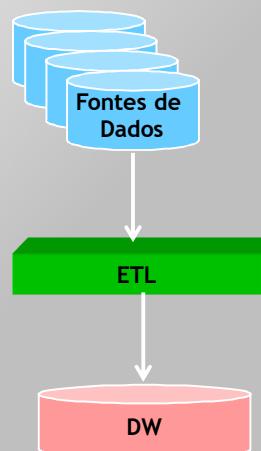
13

Arquit. - Data Marts em Conformidade



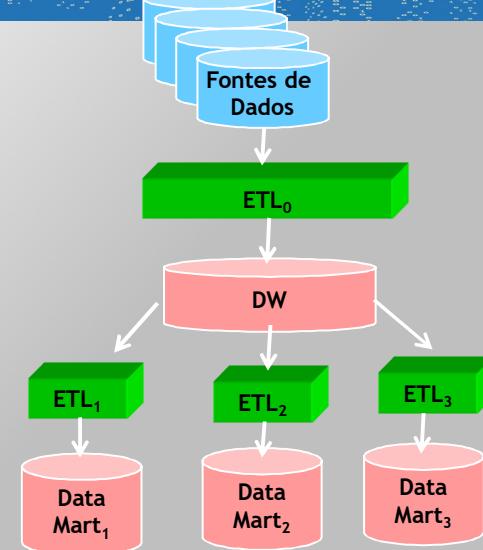
14

Arquit. - Data Warehouse Centralizado



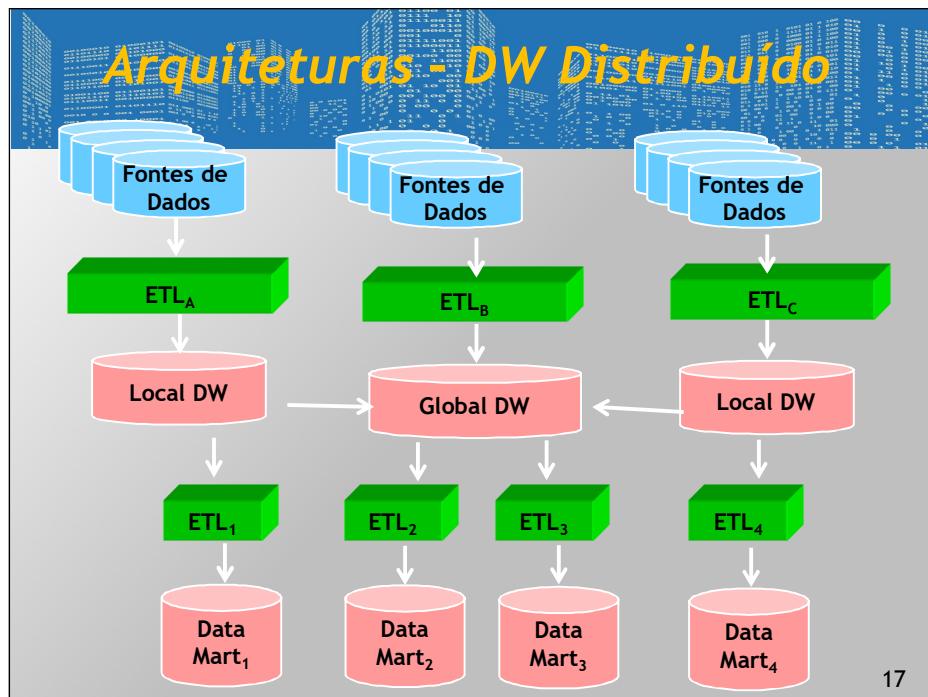
15

Arquiteturas - Hub and Spoke



16

Arquiteturas - DW Distribuído



17

Comparação Data Warehouse x Bancos de Dados Transacionais

18

Comparação

Características	BD Transacional	DW
Objetivo	Rodar o Negócio	Análise do Negócio
	Operações Diárias	Decisões Estratégicas de Longo Prazo
Tipo de Informação	Operacional	Informativo/ Analítico
Unidade de Trabalho	Consulta, Inserção, Alteração, Exclusão	Carga e Consulta
Nº Usuários	10X	X
Tipo Usuários	Operadores	Gerência
Interação Usuários	Somente Pré-Definida	Pré-Definida e Ad-Hoc

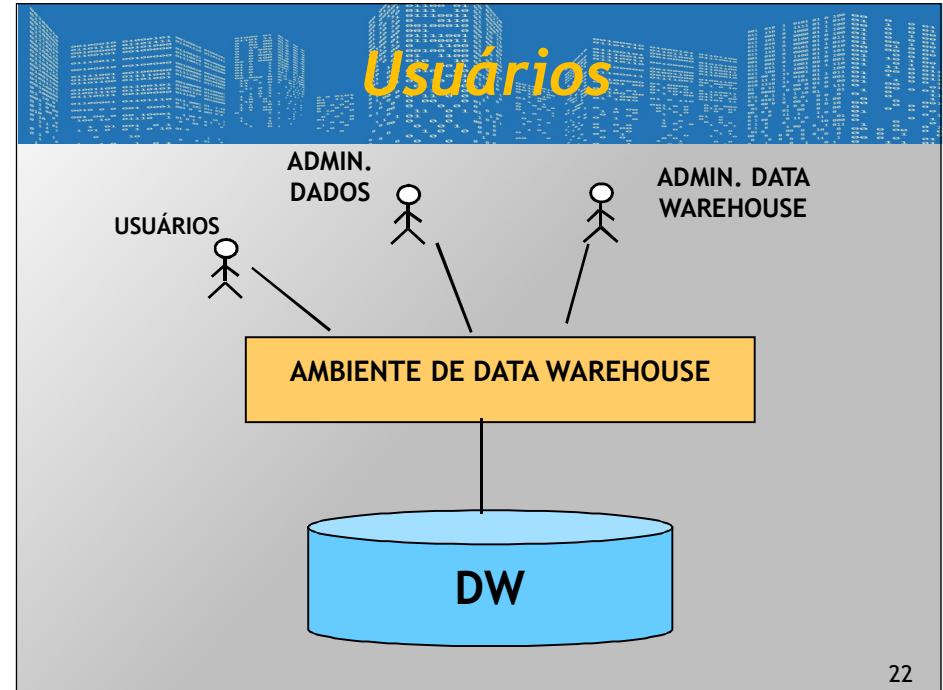
Comparação

Características	BD Transacional	DW
Volume	MB - GB	GB - TB
Granularidade	Dados Detalhados	Dados Detalhados e Resumidos
Dados	Atuais/ Operacionais	Históricos
Nº Registros/ Transação	Pequeno	Grande
Acessos	Grande Volume de Transações	Médio Volume de Análise
Utilização	Constante (alta)	Picos

20



21



22

Ambiente DW

Interação com o Gerenciador de Arquivos

- Utilizar o Sistema Operacional

Garantia de Integridade

- Impor restrições nos dados
- Verificar se violam regras de integridade

Garantia da Segurança

- Nem todo usuário do DW deve ter acesso a todo conteúdo/operações do DW

Recuperação e Backup

- Falha: mecanismos de recuperação → restaurar DW à situação original ou
- Facilidades de recuperação - cópias de segurança

Controle de Concorrência

- Controle da interação - usuários concorrentes → não violar consistência de dados

23

Ambiente DW

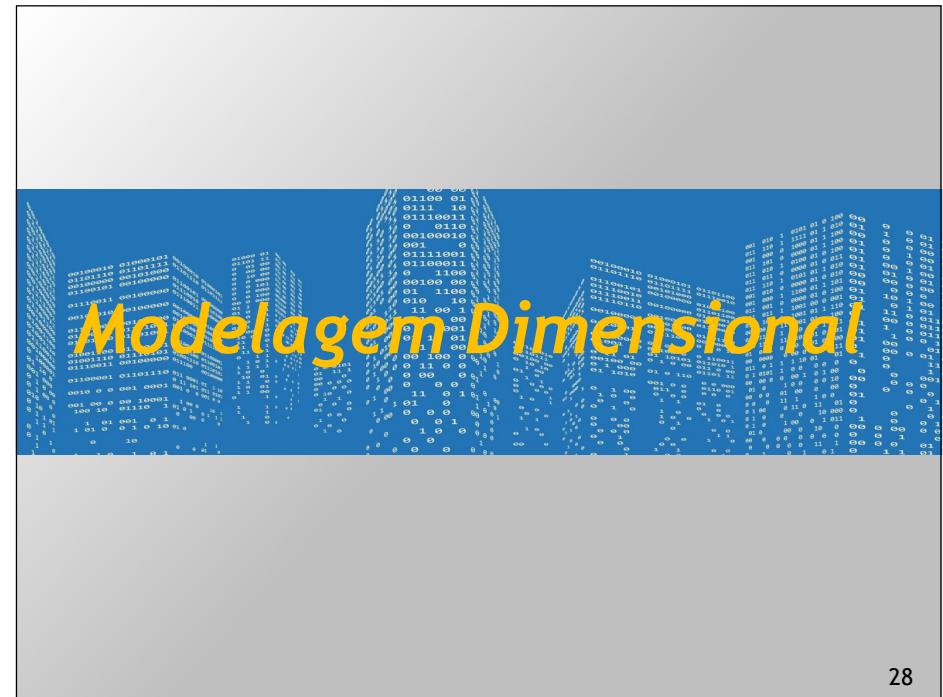
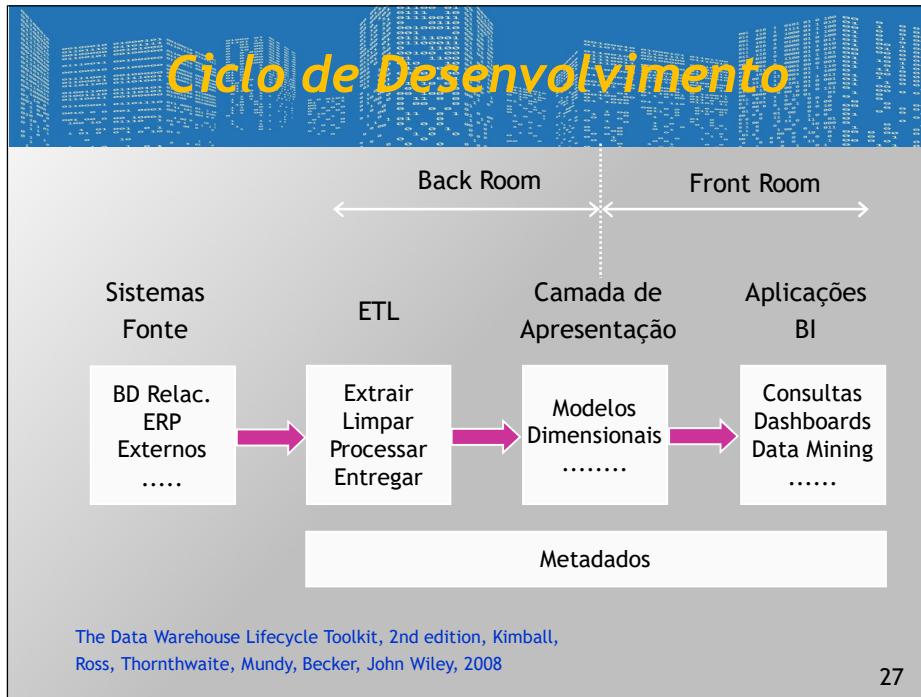
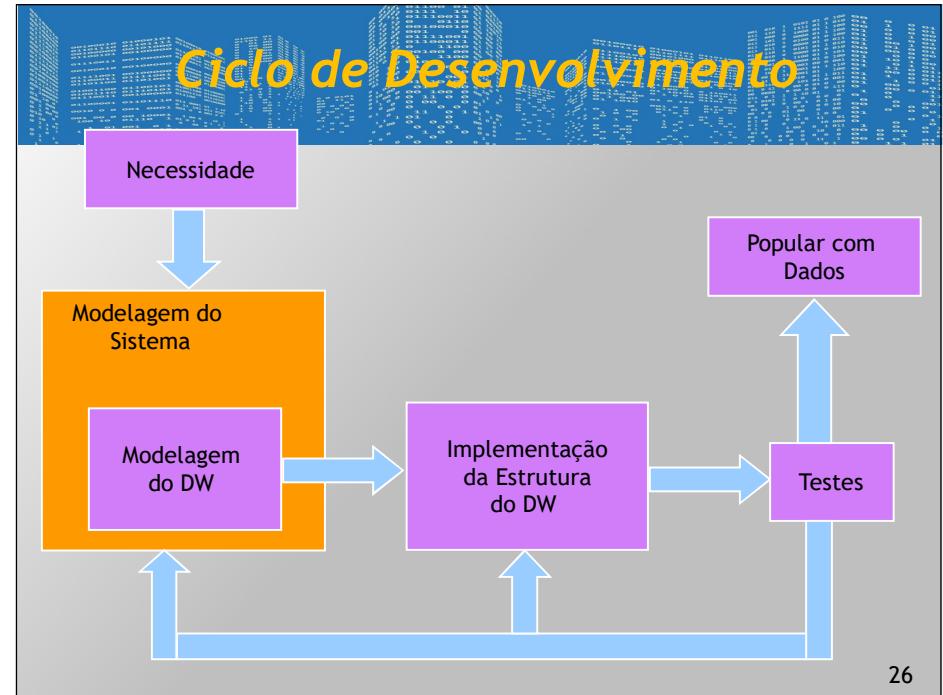
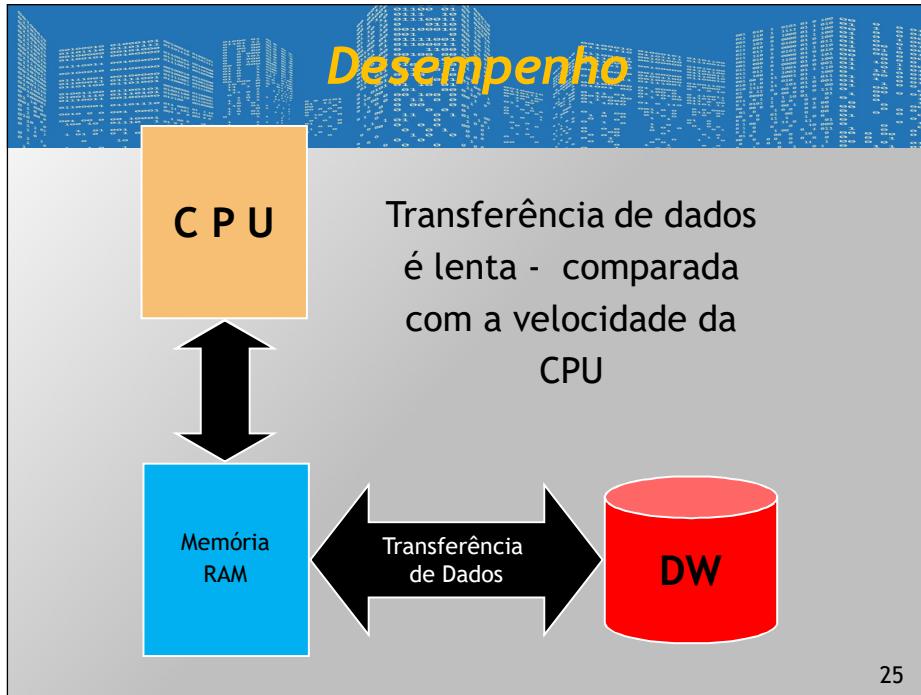
Recuperação e Backup

- Falha: mecanismos de recuperação → restaurar DW à situação original ou
- Facilidades de recuperação - cópias de segurança

Controle de Concorrência

- Controle da interação - usuários concorrentes → não violar consistência de dados

24



Modelagem do DW

Estrutura de dados

- Medições e dimensões

Medições

- Dados numéricos: **Tabela Fato**

Dimensões

- Parâmetros do negócio
- Tabelas satélites vinculadas à Tabela Fato central: **Tabelas Dimensão**

29

Modelagem Dimensional (Star Join Schema)

Diagrama semelhante a uma estrela

- Tabela ‘grande’ no centro rodeada por tabelas “auxiliares”
- Cada tabela “auxiliar” (**Tab. Dimensão**) tem relacionamento 1 x N com a tabela central (**Tabela Fato**)

31

Modelagem do DW

Modelo adequado segundo visão dos usuários

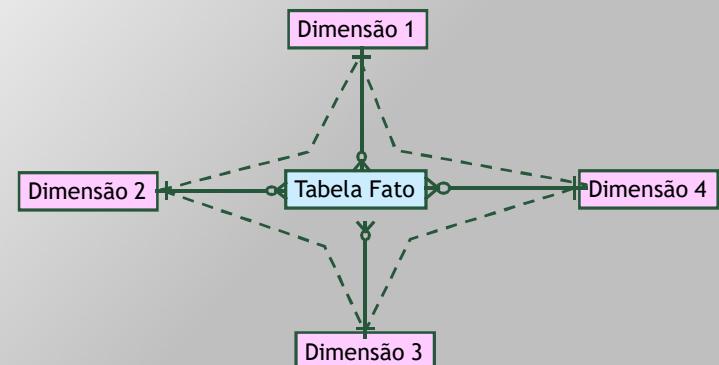
Modelo Dimensional

Estrutura simples

Não normalizado

30

Modelagem Dimensional



32

Modelagem Dimensional

Tabelas Dimensão

Chave Dimensão 1
Atributo 1.1
Atributo 1.2
.....
Atributo 1.n

Chave Dimensão 2
Atributo 2.1
Atributo 2.2
.....
Atributo 2.n

Tabela Fato

Chave Dimensão 1
Chave Dimensão 2
Chave Dimensão 3
Chave Dimensão 4

Fato 1
Fato 2
.....
Fato n

Tabelas Dimensão

Chave Dimensão 3
Atributo 3.1
Atributo 3.2
.....
Atributo 3.n

Chave Dimensão 4
Atributo 4.1
Atributo 4.2
.....
Atributo 4.n

33

Modelagem Dimensional

Tabela Fato

Chave Dimensão 1
Chave Dimensão 2
Chave Dimensão 3
Chave Dimensão 4

Fato 1
Fato 2
.....
Fato n

Valores numéricos (medidas)

Cada valor: interseção de todas dimensões

Explícita: não há dados em todas interseções

Nível de detalhe da tabela fato → granularidade

Indicadores importantes para uma área de negócios

34

Modelagem Dimensional

Tabelas Dimensão

Chave Dimensão 3
Atributo 3.1
Atributo 3.2
.....
Atributo 3.n

Descrições (textuais) do negócio

Definem propriedades da dimensão

Campo numérico: fato ou atributo ?

- Se varia a cada amostragem → fato
- Se é uma descrição praticamente constante de um item → atributo

Chaves Substitutas

35

Modelagem Dimensional

Dimensões desnormalizadas

Modelo simples

Diminui necessidade de junções → melhora no desempenho

Expansão simplificada do modelo do DW

36

Modelagem Dimensional

Dimensão Tempo

Chave Tempo
dia da semana
mês
trimestre
ano

Fato Vendas

Chave Tempo
Chave Produto
Chave Loja
Chave Cliente
valor vendas
unidade vendidas
custo

Dimensão Produto

Chave Produto
descrição
marca
categoria

Dimensão Cliente

Chave Cliente
Cidade
Estado
.....

Dimensão Loja

Chave Loja
nome loja
endereço
tipo loja

37

Modelagem Dimensional

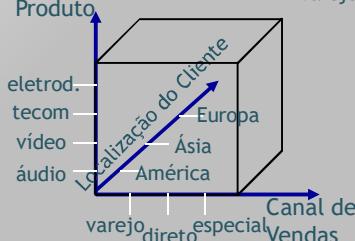
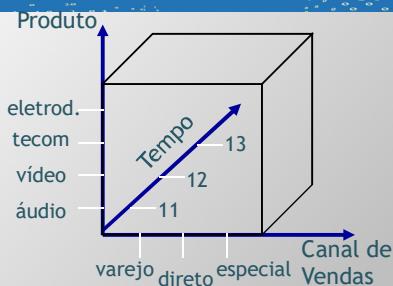
Vendas no 1º Trimestre/2019

Marca	Valor Vendas	Unidades Vendidas
Axon	700,00	200
Frami	850,00	221
Triz	620,00	170
Zigzag	1.200,00	340

```
select p.marca, sum (f.valor vendas),
       sum (f.unidades vendidas)
  from Vendas f, Produto p, Tempo t
 where f.chave produto = p.chave produto AND
       f.chave tempo = t.chave tempo AND
       t.trimestre = '1T2019'
 group by p.marca
 order by p.marca
```

38

Modelagem Dimensional



39

Modelagem Dimensional

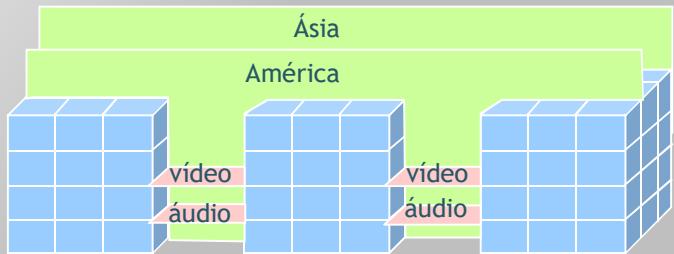
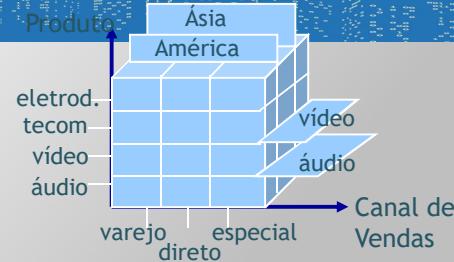
❖ **Cubo de Dados:** denominação de uma estrutura dimensional produzida por uma consulta

❖ **Dimensões originais:** produto, loja e tempo

Cube. Another term for a fact table. It can represent “n” dimensions, rather than just three (as may be implied by the name).

40

Modelagem Dimensional

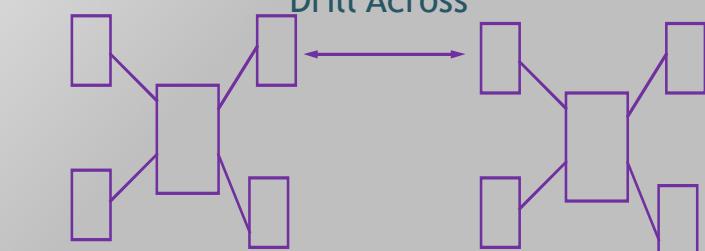


41

Roll Up/Drill Down/ Drill Across



Drill Across



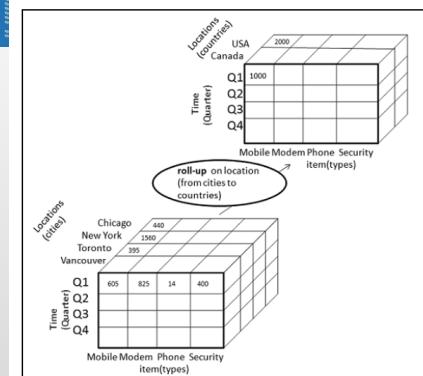
42

Cubo

		Mumbai	New Delhi	Gurgaon	
		986	567	875	908
Locations (cities)		786	85	987	
Time (Quarter)	Q1	788	987	765	
Time (Quarter)	Q2	678	654	987	
Time (Quarter)	Q3	899	875	190	
Time (Quarter)	Q4	787	969	908	
		Mouse	Mobile	Modem	item(types)

43

Roll Up/Drill Down



		USA	Canada	
		2000	1000	2000
Locations (countries)		Q1	Q2	Q3
		Mobile	Modem	Phone
		Security	item(types)	

		USA	Canada	
		2000	1000	2000
Locations (countries)		Q1	Q2	Q3
		Mobile	Modem	Phone
		Security	item(types)	

		USA	Canada	
		2000	1000	2000
Locations (countries)		Q1	Q2	Q3
		Mobile	Modem	Phone
		Security	item(types)	

		Chicago	New York	Toronto	Vancouver	
		440	1560	395		395
Locations (countries)		Q1	Q2	Q3	Q4	
		Mobile	Modem	Phone	Security	item(types)
		Security	item(types)			

		Chicago	New York	Toronto	Vancouver	
		440	1560	395		395
Locations (countries)		Q1	Q2	Q3	Q4	
		Mobile	Modem	Phone	Security	item(types)
		Security	item(types)			

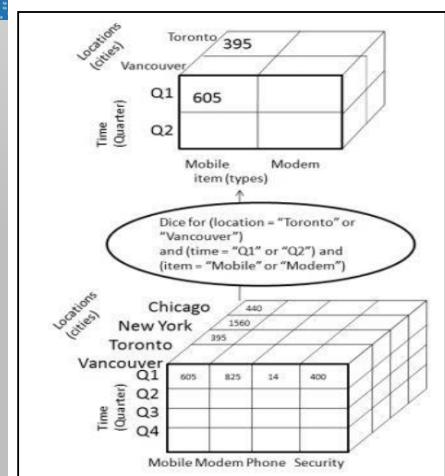
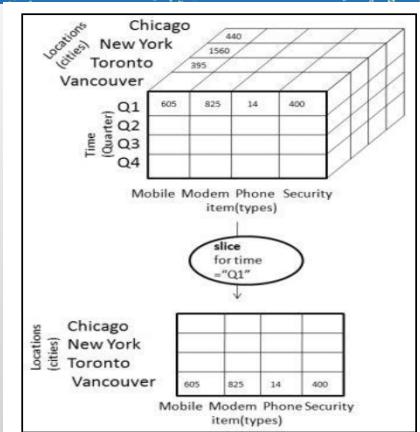
		Chicago	New York	Toronto	Vancouver	
		440	1560	395		395
Locations (countries)		Q1	Q2	Q3	Q4	
		Mobile	Modem	Phone	Security	item(types)
		Security	item(types)			

		Chicago	New York	Toronto	Vancouver	
		440	1560	395		395
Locations (countries)		Q1	Q2	Q3	Q4	
		Mobile	Modem	Phone	Security	item(types)
		Security	item(types)			

		Chicago	New York	Toronto	Vancouver	
		440	1560	395		395
Locations (countries)		Q1	Q2	Q3	Q4	
		Mobile	Modem	Phone	Security	item(types)
		Security	item(types)			

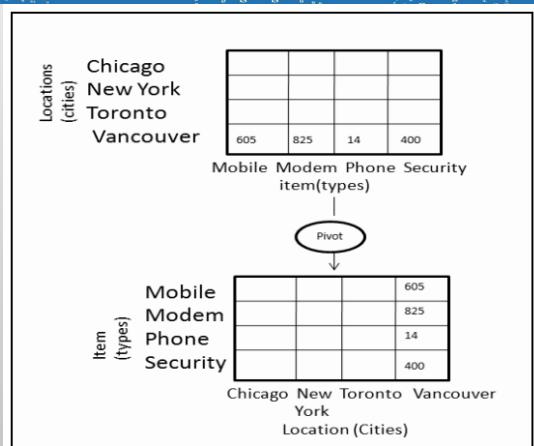
		Chicago	New York	Toronto	Vancouver	
		440	1560	395		395
Locations (countries)		Q1	Q2	Q3	Q4	
		Mobile	Modem	Phone	Security	item(types)
		Security	item(types)			

Slice and Dice



45

Pivot / Rotation



46

Modelagem Dimensional

❖ Drill Down

Marca	Tam.emb.	Vendas
BXO	2	50
BXO	3	110
BXO	6	75

Marca	Tam.emb.	Cor	Vendas
BXO	2	branco	8
	2	marrom	5
	2	verde	37
	3	branco	22
	3	verde	88
	6	branco	14
	6	rosa	12
	6	marrom	4
	6	verde	45

47

MDX (Multi-Dimensional eXpressions)

48

- ❖ MDX (Multi-Dimensional eXpressions):
linguagem de consulta utilizada em bases
de dados multidimensionais
- ❖ Desenvolvida pela Microsoft e
disponibilizada no Analysis Services 7.0 -
1998

- ❖ Utiliza
 - Modelo dimensional com suas hierarquias
 - Membros: valores das tabelas dimensão
 - Medidas: valores presentes nas tabelas fato

- ❖ Formato para acessar um membro
[Dimensão].[Hierarquia de Atributo].[Membro
da Hierarquia] → eixo A
- ❖ Formato para acessar uma medida
Measures.[Membro numérico da
Tabela Fato] → eixo B
- ❖ Por exemplo
[Cliente].[País].[Canadá] → eixo A
Measures.[Vendas] → eixo B

- ❖ Consulta básica

```
SELECT eixoA1, eixoA2, ..., eixoAn ON COLUMNS
      eixoB1, eixoB2, ..., eixoBn ON ROWS
  FROM modelo dimensional
 WHERE condição
```

MDX

SELECT Measures.[Total de Vendas] on COLUMNS,
{[Cliente].[País].[Brasil],
[Cliente].[País].[Canadá],
[Cliente].[País].[Espanha]} on ROWS
FROM [Vendas]

País	Total de Vendas
Brasil	3.552,00
Canadá	5.994,00
Espanha	2.777,00

53

MDX - Keywords

- ❖ **Members:** obter todos os membros de um nível da hierarquia

SELECT Measures.[Vendas] on COLUMNS,
[Cliente].[País].members on ROWS
FROM [Vendas]

País	Total de Vendas
Brasil	3.552,00
Canadá	5.994,00
Espanha	2.777,00
China	1.222,00
USA	6.897,00
....

4

MDX - Keywords

- ❖ **Children:** obter todos os membros filhos de um membro

SELECT Measures.[Vendas] on COLUMNS,
[Cliente].[País].[Brasil].children on ROWS
FROM [Vendas]

Brasil	Total de Vendas
SP	3.552,00
RJ	5.994,00
MG	2.777,00
....

55

Granularidade

56

Granularidade

❖ Nível de detalhe/resumo nos dados

Maior Detalhamento
Maior Granularidade



Menor Detalhamento
Menor Granularidade

❖ Afeta diretamente

- Volume de dados
- Tipos de consultas que podem ser atendidas

❖ Definição importantíssima

57

Granularidade

alto nível de detalhes
 detalhes de cada chamada telefônica de cada cliente

João ligou para o DETRAN no dia 05/04/20 ?

pode ser respondida

Porém a procura por um único registro é um fato raro

Em média, quantas ligações são feitas por mês ?

Possível - nº imenso de registros

baixo nível de detalhes
 resumo das chamadas telefônicas de cada cliente por mês

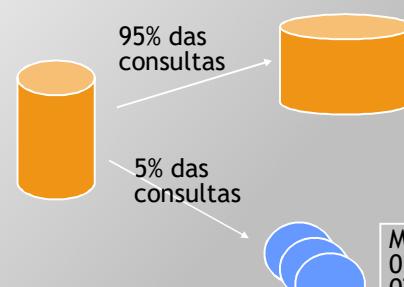
não pode ser respondida

Janeiro
Maria da Silva
Nº de chamadas: 51
.....

Nível Duplo de Granularidade

❖ Empresa possui dados em demasia no DW

- Dois ou mais níveis de granularidade



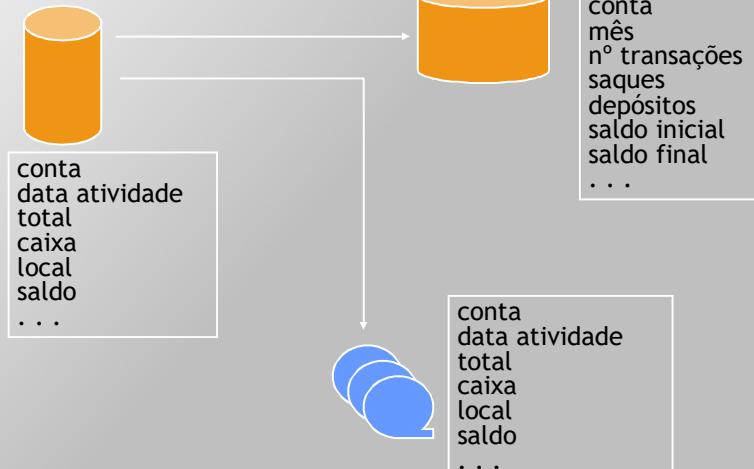
Maria da Silva
05/04/99 - DDD - Campinas ...
07/04/99 - DDI - França ...
.....

59

60

Granularidade

Exemplo: banco



61

Granularidade

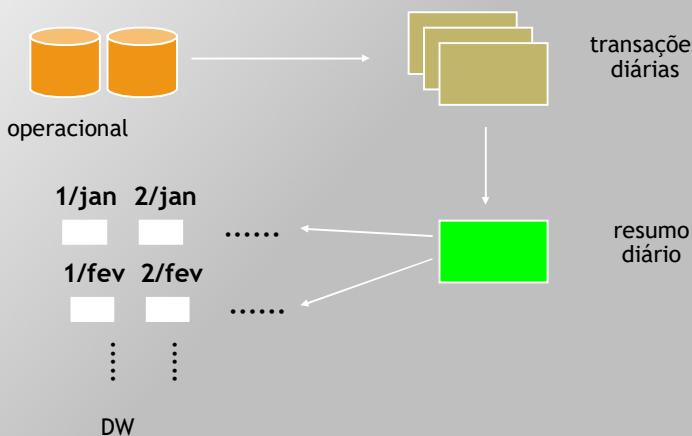
❖ Definição dos níveis de granularidade:

- Processo iterativo com os usuários
- Formas de diminuir a granularidade
 - ✓ Resumos
 - ✓ Médias
 - ✓ Valores limite
 - ✓ . . .

62

Estruturação dos Dados

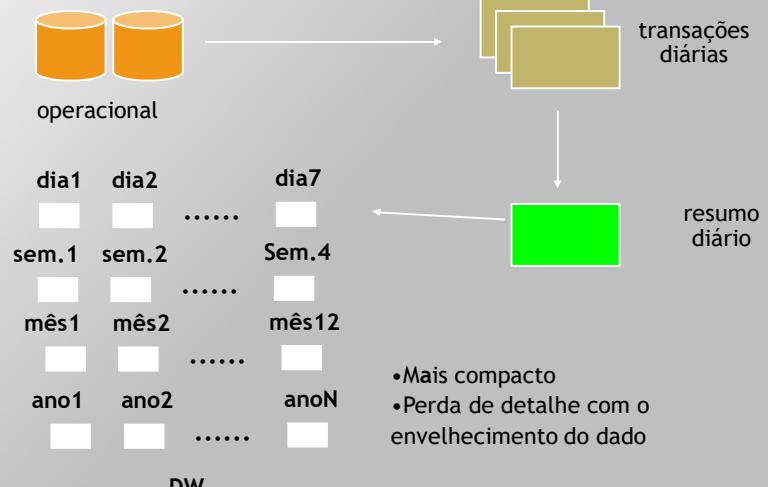
❖ Cumulativa



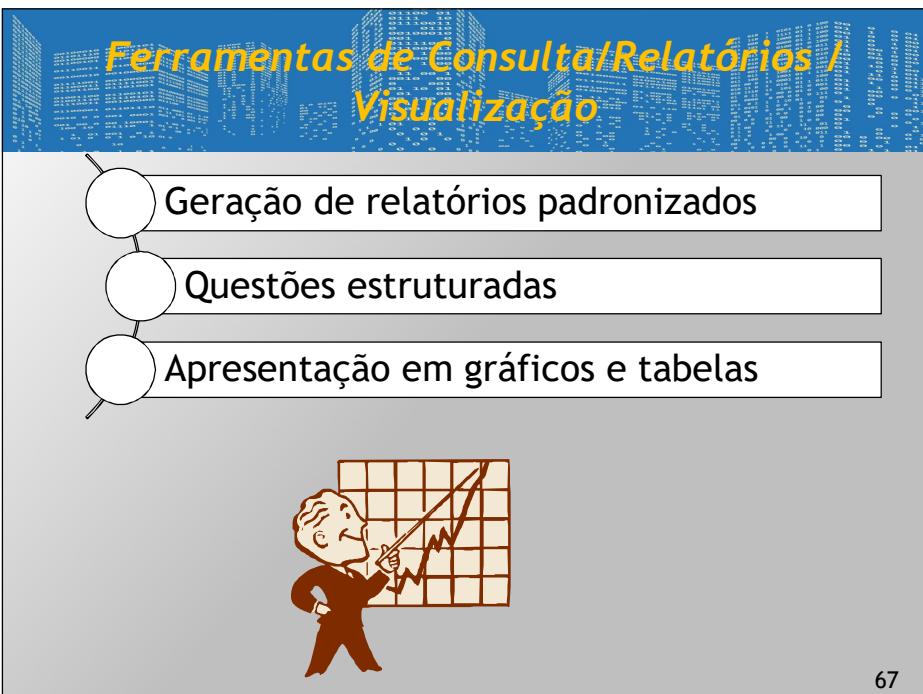
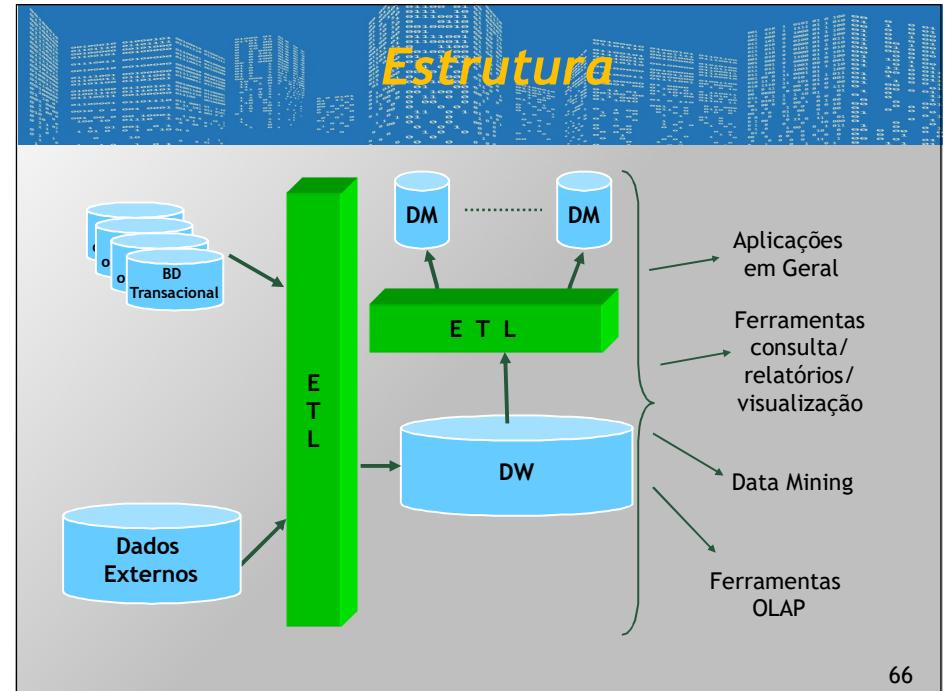
63

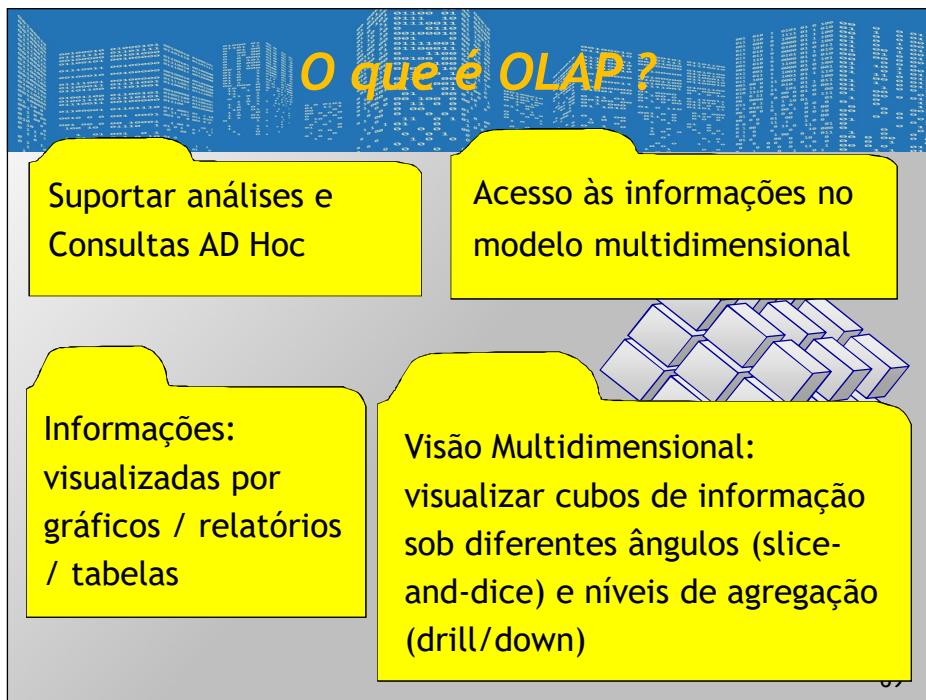
Estruturação dos Dados

❖ Resumo Rotativo



64





70



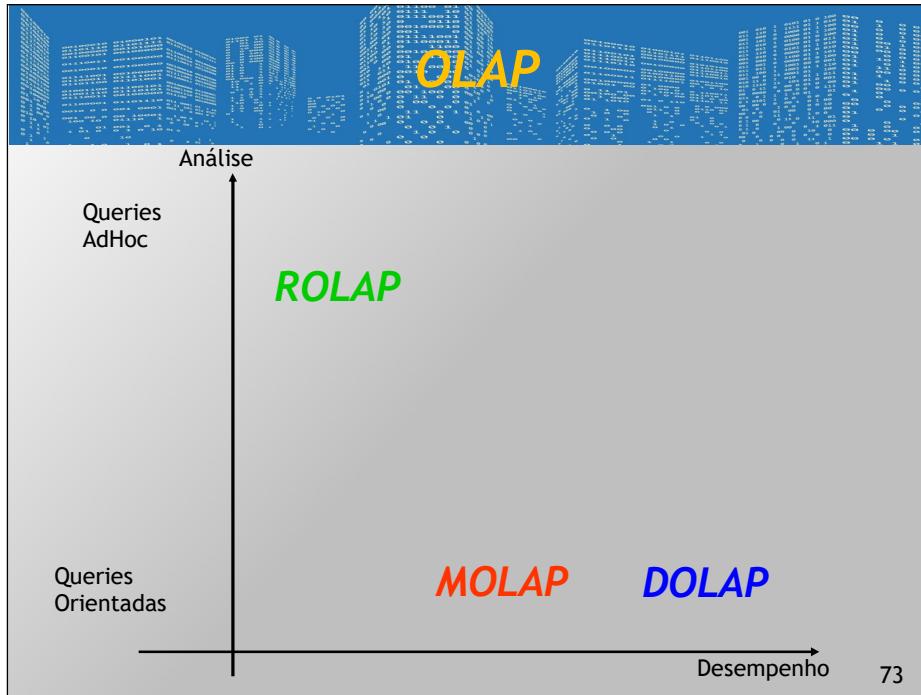
70



71



72



Outros Tipos de OLAP

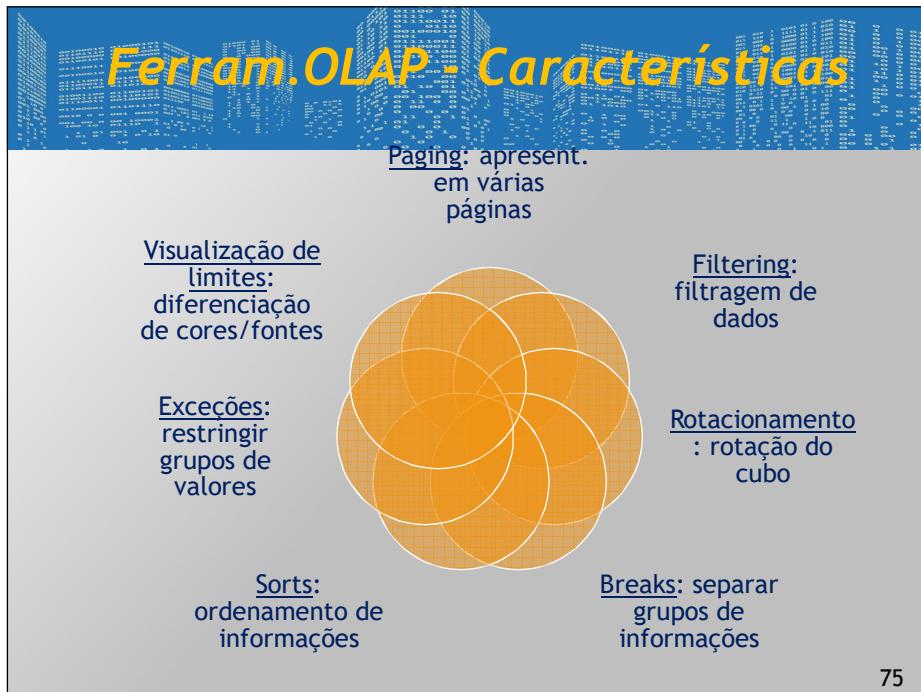
WOLAP (Web OLAP): acesso apenas por navegadores

Mobile OLAP: aprimorar acesso por dispositivos móveis

SOLAP (Spatial OLAP): integração com GIS

HTAP (Hybrid Transactional/Analytical Processing): in-memory data store

74



Mercado OLAP - Pesquisa

Mercado de bilhões de US\$

Alta taxa de “shelfware”: compra de 2 produtos extras a cada 3 utilizados

50%: consideráveis mudanças organizacionais

Implementações OLAP: mais problemas com políticas da empresa do que com os produtos em si

“The OLAP Survey” (questionário de 644 pessoas de 32 países)

76

Mercado OLAP - Pesquisa

Problema mais comum: baixo desempenho das consultas

Poucas reclamações de confiabilidade/segurança

Altos preços não necessariamente garantem melhores resultados

77

Mercado OLAP - Pesquisa

Definir necessidades de negócio antes de contatar fornecedores

- Descobrir o que os usuários realmente precisam, não o que eles dizem precisar
- Envolver usuários finais em todos os estágios
- Volume de dados (atual e futuro)

78



79

By Jeffrey Walker Last updated on 10.01.2018

Top 5 data warehouses on the market today

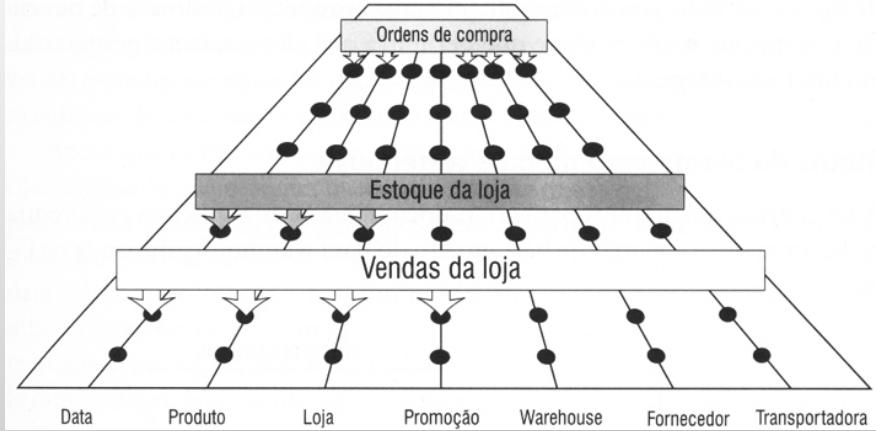


80

Matriz de Barramento do Data Warehouse

81

Matriz de Barramento do DW



82

Matriz de Barramento do DW

	Dime nsão 1	Dime nsão 2	Dime nsão 3	Dime nsão 4	Dime nsão 5	...
Área Negócio 1	X		X	X	X	
Área Negócio 2	X		X			
Área Negócio 3	X	X		X	X	
Área Negócio 4		X	X	X		
.....						

83

Matriz de Barramento do DW

Dimensão Área Negócio	Da ta	Pro du to	Loj a	Pro mo ção	De pó si to	Forn ece dor	Com tra to	Trans port.
Vendas Varejo	X	X	X	X				
Estoque Varejo	X	X	X					
Entregas Varejo	X	X	X					
Estoque Depósito	X	X			X	X		
Entregas Depósito	X	X			X	X		
Ordens Compra	X	X			X	X	X	X

84

Matriz de Barramento do DW

- Mesmos significados
- Eventual agrupamento/replicação

Dimensões em Conformidade

Modificações no Esquema Estrela

- Novos atributos de dimensão
- Novas dimensões
- Novos fatos
- Mudanças de granularidade

85

Partial Star

Múltiplas tabelas para cada dimensão e/ou para os fatos

Separação/criação em função do nível de sumarização

87

Variantes do Modelo Dimensional Partial Star

86

Star Completo

Dimensão Localização

Chave Localização
Descrição
...
Cidade
...
Estado
...
Região
...

Dimensão
Taxa Juros

Chave Tx Juros
Descrição
Taxa Agregada

Fato Vendas

Chave Localização
Chave Produto
Chave Tempo
Chave Tx Juros
Quantidade
Custo
Venda
Valor
Nº Parcelas

Dimensão Produto

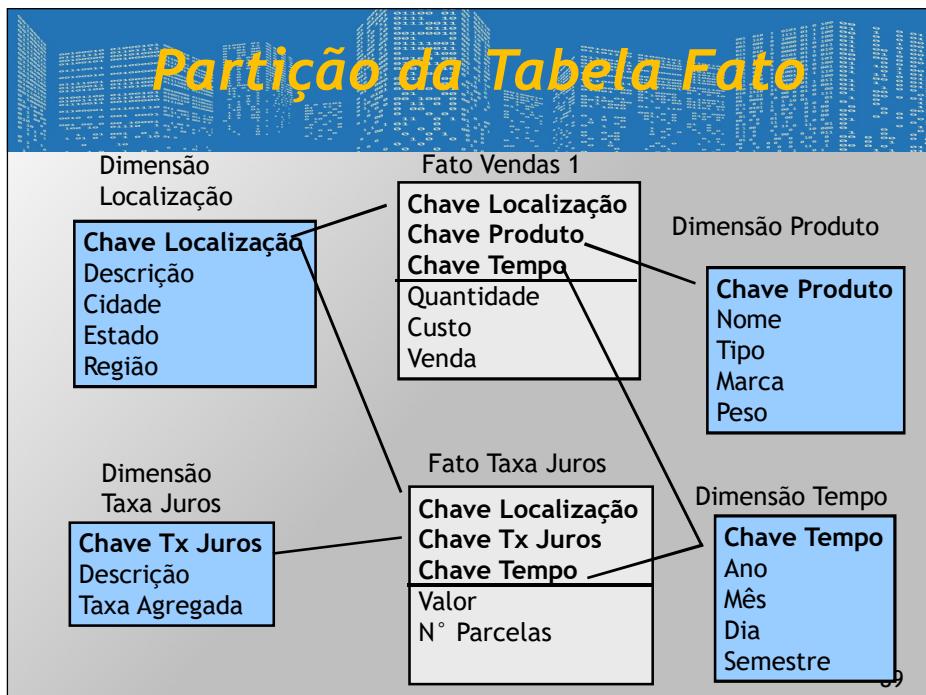
Chave Produto
Nome
Tipo
Marca
Peso

Dimensão Tempo

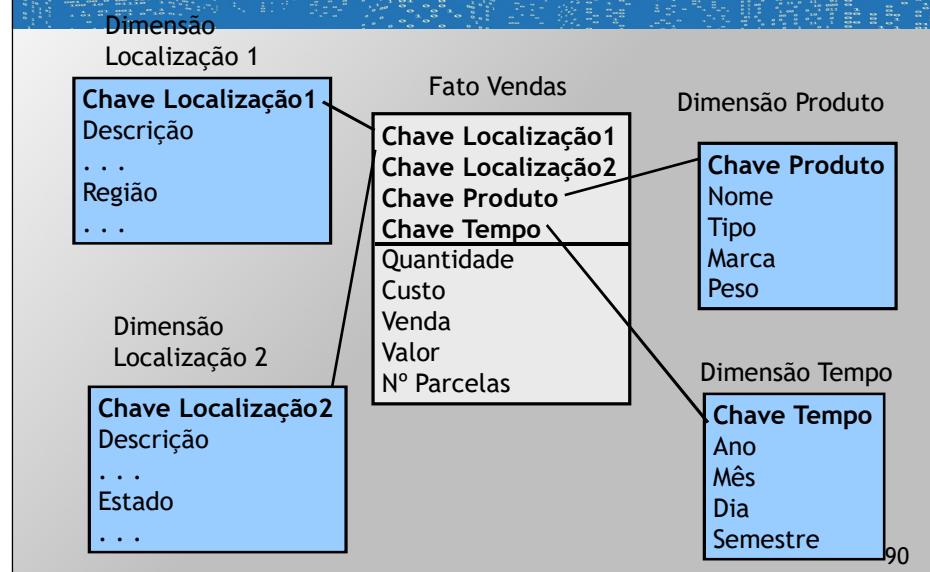
Chave Tempo
Ano
Mês
Dia
Semestre

88

Partição da Tabela Fato



Partição das Tabelas Dimensão

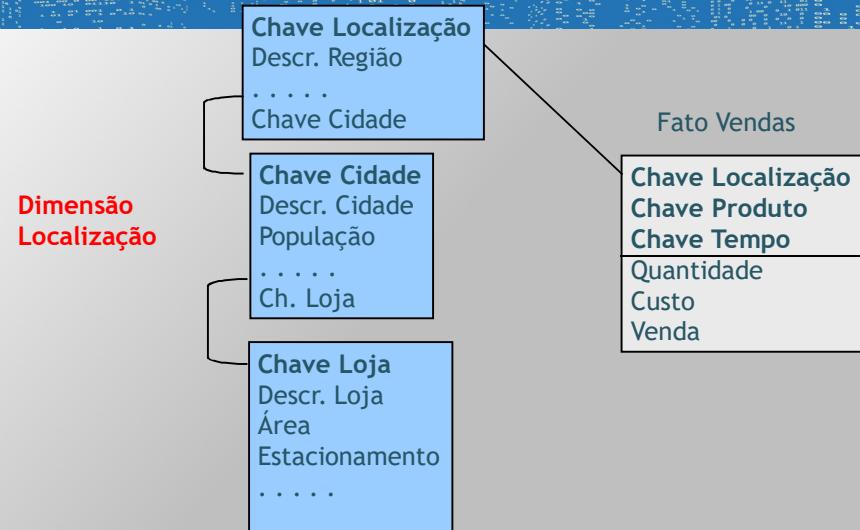


Variantes do Modelo Dimensional Snowflake

Esquema Snowflake

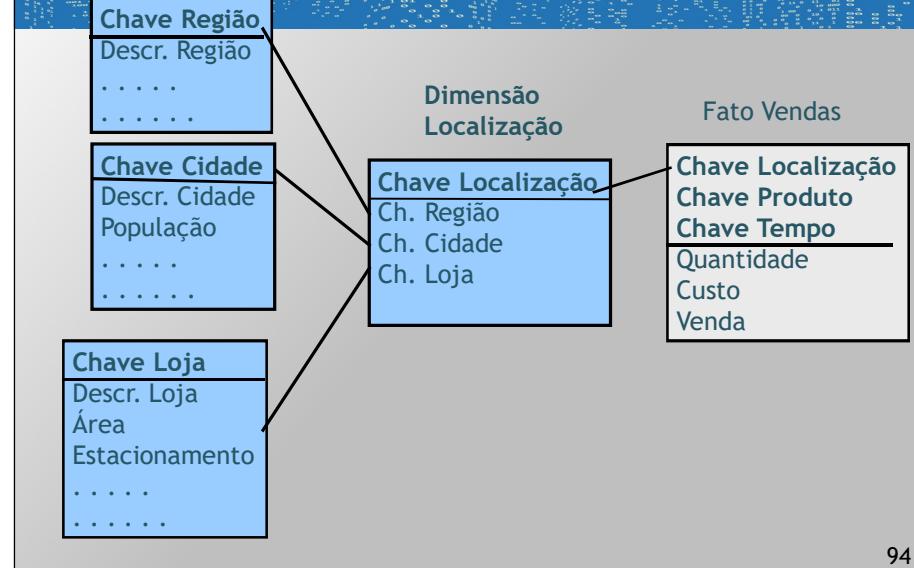
- Normalizar tabelas Dimensão
- Desvantagens: complexidade maior e menos intuitivo
- Vantagem: espaço de armazenamento
- Snowflake parcial: normalizar algumas dimensões
- Duas variações: lookup e chain

Snowflake Chain



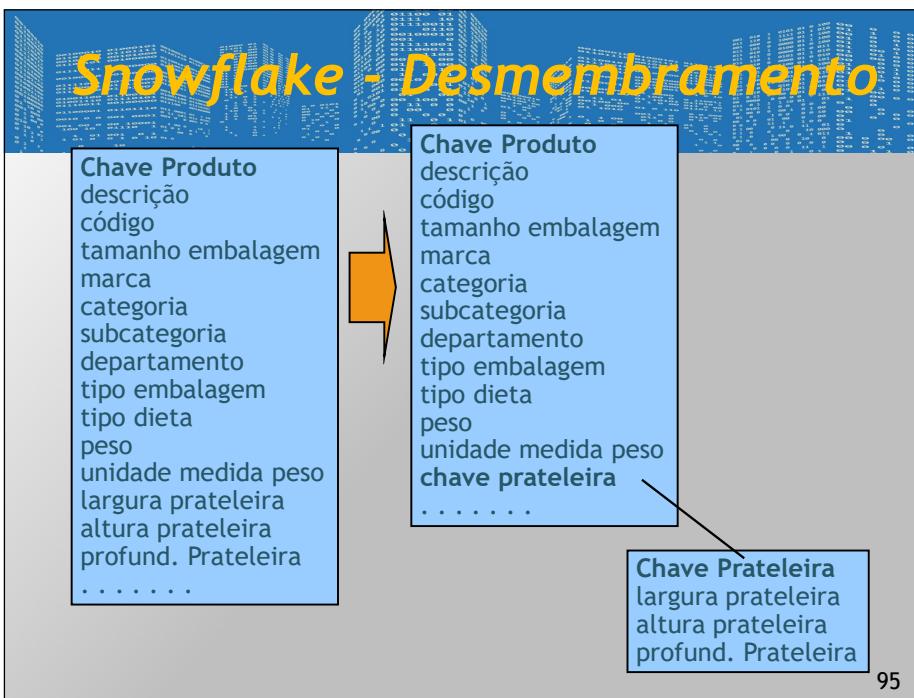
93

Snowflake Lookup



94

Snowflake - Desmembramento



95

Variantes do Modelo Dimensional Minidimensões

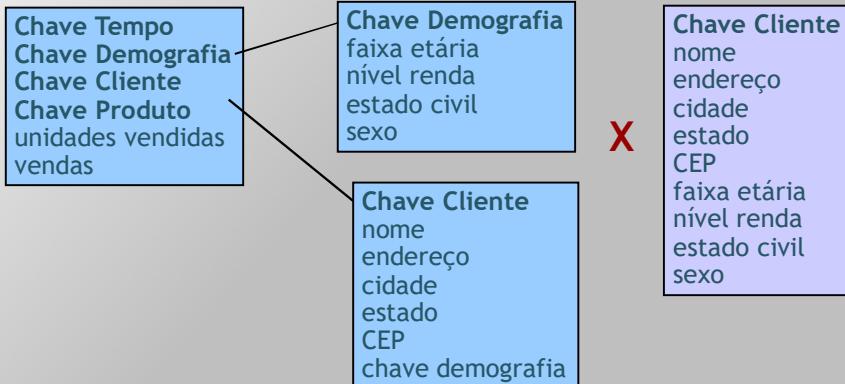


96

Modelagem Dimensional

❖ Minidimensões - nº pequeno de atributos

Fato Vendas



97

Modelagem Dimensional Volume de Dados

98

Aplicação - Supermercado

Rede de lojas (300)

Cada loja: 30.000 produtos

Reduções temporárias de preços - promoções



99

Processo a ser modelado

- Movimento diário de item - quais produtos estão sendo vendidos, em que lojas, a que preço e em que dias

Definição do grão da tabela fato

- Movimento diário, por item, por loja, por promoção

100

Aplicação - Supermercado

Dimensão Tempo

Chave Tempo
atributos tempo

Dimensão Promoção

Chave Promoção
atributos promoção

Dimensão Produto

Chave Produto
atributos produto

Dimensão Loja

Chave Loja
atributos loja

Fato Vendas

Chave Tempo
Chave Produto
Chave Loja
Chave Promoção
Fatos

101

Aplicação - Supermercado

❖ Medições para tabela fato

Fato Vendas

Chave Tempo
Chave Produto
Chave Loja
Chave Promoção
valor vendas
unidades vendidas
custo
número clientes



❖ Tabela fato: muito maior que tabelas dimensão

102

Aplicação - Supermercado

Dimensão Tempo

Chave Tempo
dia da semana
nº do dia no mês
nº dias corridos
nº semana no ano
nº semanas corridas
mês
nº meses corridos
trimestre
período fiscal
indicador feriado
.....

Dimensão Produto

Chave Produto
descrição
código
tamanho embalagem
marca
categoria
subcategoria
departamento
tipo embalagem
peso
unidade medida peso
largura prateleira
altura prateleira
profund. prateleira
.....

Dimensão Loja

Chave Loja
nome loja
número loja
endereço
cidade
estado
CEP
gerente
telefone
tipo planta
data abertura
área útil total
área padaria
área açougue
.....

Dimensão Promoção

Chave Promoção
nome promoção
tipo de preço
tipo de anúncio
tipo de display
nome mídia anúncio
custo promoção
data início promoção
data término promoção

103

Aplicação - Supermercado

2 anos → 750 dias, 300 lojas, vendas diárias

30.000 produtos em cada loja

- 3.000 vendidos todos os dias em cada loja

Um item tem apenas uma promoção por dia

$750 \times 300 \times 3.000 \times 1 = 657$ milhões de reg.

Tabela Fato: 8 campos x 4 bytes = 32 bytes → 21 GB

104

- ❖ Dimensão Produto
- ❖ 30.000 itens
- ❖ 100 atributos
- ❖ 10 bytes
- ❖ $30.000 \times 100 \times 10 = 30.000.000$
- ❖ 30 MB

105

Dimensões de Modificação Lenta

Dimensões de Modificação Lenta

Dados das tabelas dimensão podem sofrer modificações (estado civil, endereço, filhos, nomes de dptos, . . .)



107

Substituição de valores antigos

- Perda da capacidade de análise histórica
- Usado quando o valor antigo perde significado

Criação de novos campos para indicar o estado atual do objeto

Criação de novos registros com os novos valores

- Necessidade de usar chaves generalizadas
- Divisão do histórico considerando os diversos registros
- Mantém histórico

108

Dimensões de Modificação Lenta

Tabela Original

ID	Empresa
1	ABC
2	XYZ

Tabela com subst. de valores

ID	Empresa
1	Delta
2	Beta

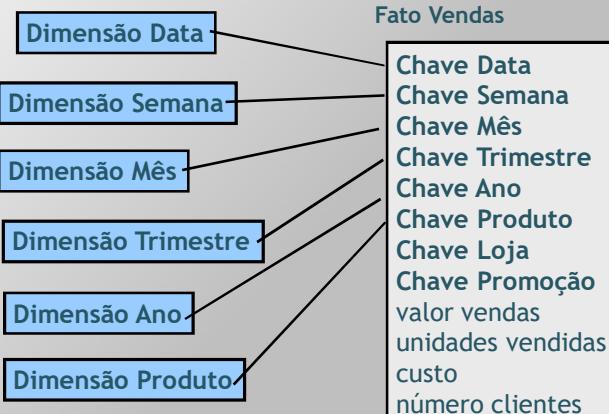
Tabela com novos campos

ID	Empresa 1	Data 1	Empresa 2	Data 2
1	ABC	03/06/10	Delta	06/05/18
2	XYZ	03/06/10	Beta	06/05/18

Tabela com novos registros

ID	Empresa	Data
1	ABC	03/06/10
1A	Delta	06/05/18
1B	New	11/11/20
2	XYZ	03/06/10
2A	Beta	06/05/18

Excesso de Dimensões



Procurar as combinações adequadas



Variantes do Modelo Dimensional Cesta de Mercado

Cesta de Mercado

Fato Vendas

Chave Tempo
Chave Produto A
Chave Produto B
Chave Loja
Chave Promoção
Qtdade Vendida Prod. A
Qtdade Vendida Prod. B
Valor Venda Prod. A
Valor Venda Prod. B

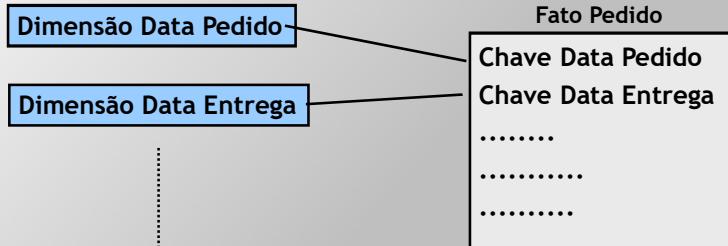
Excesso de Registros: Remoção Progressiva

113

Variantes do Modelo Dimensional
Representação de Papeis

114

Representação de Papeis



115

Variantes do Modelo Dimensional
Dimensão de Degeneração

116

Dimensão de Degeneração

Dimensão vazia

Informação está em outras dimensões

Fato Vendas

Chave Data
Chave Produto
Chave Loja
Chave Promoção
Nº Transação Comercial (DD)
.....
.....

117

Modelagem Dimensional Volume de Dados

118

Aplicação - Bancos

Chave Conta
Nome Titular
Nome 2º Titular
Endereço
Cidade
Estado
CEP
Data Abertura
Sexo
Estado Civil

Chave Conta
Chave Agência
Chave Produto
Chave Status
Chave Tempo
Saldo
Operação

Chave Agência
Nome Agência
Endereço
Cidade
Estado
CEP
Tipo Agência

Chave Produto
Descrição
Tipo
Categoria

Chave Status
Descrição
Motivo
Indic. Conta Nova
Indic. Conta Encerrada

Chave Tempo
Mês
Ano
Trimestre

119

Aplicação - Bancos

Dimensões heterogêneas: diversos produtos com atributos específicos

Estender tabela dimensão produto e tabela fato para acomodar atributos → muitos campos vazios

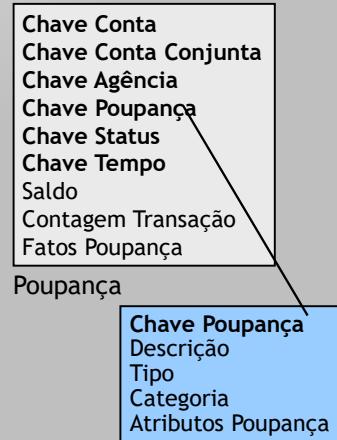
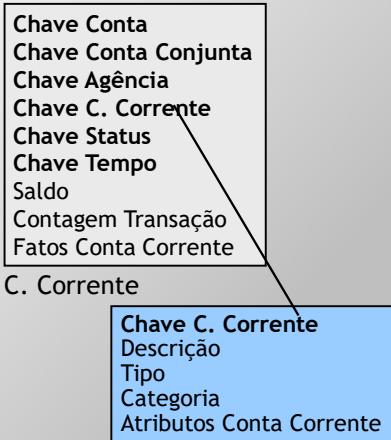
Chave Conta
Chave Conta Conjunta
Chave Agência
Chave Produto
Chave Status
Chave Tempo
Saldo
Contagem Transação
Fatos Conta Corrente
Fatos Poupança
.....

Chave Produto
Descrição
Tipo
Categoria
Atributo Conta Corrente
Atributo Poupança
.....

120

Aplicação - Bancos

- ❖ Tabelas fato e dimensão específicas para cada produto

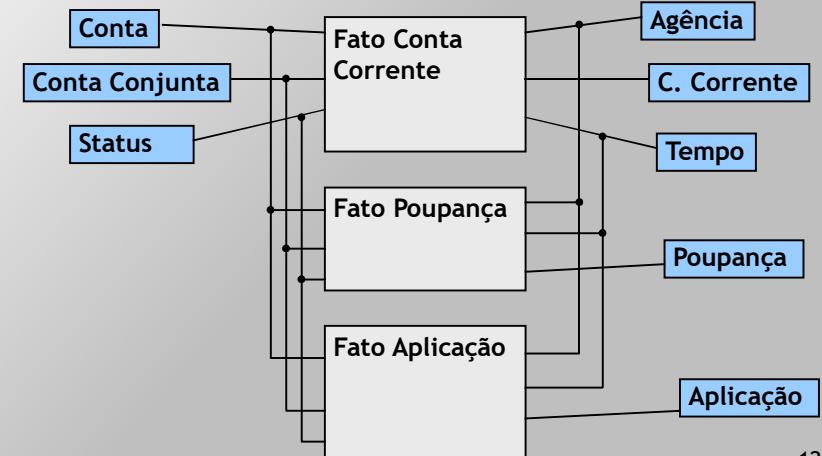


121

Aplicação - Bancos

- ❖ Tabelas Multifato - Fact Partitioning

- Várias tabelas fato com dimensões em comum



122

Aplicação - Bancos

Dimensão conta - 12 milhões de clientes

3 anos → 36 instantâneos mensais

agência, produto, status → 1/conta/mês

$12.000.000 \times 36 = 432$ milhões reg.

Central → 6 campos chave + 2 fatos →
 $432.000.000 \times 8 \times 4$ bytes = 13,8 GB

Específico → 6 campos chave + 6 fatos →
 $432.000.000 \times 12 \times 4$ bytes = 20,7 GB

123

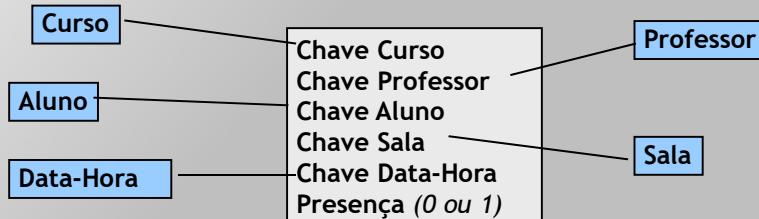
Variantes do Modelo Dimensional Tabela Fato sem Fatos

124

Tabela Fato Sem Fatos

Sem fatos mensuráveis

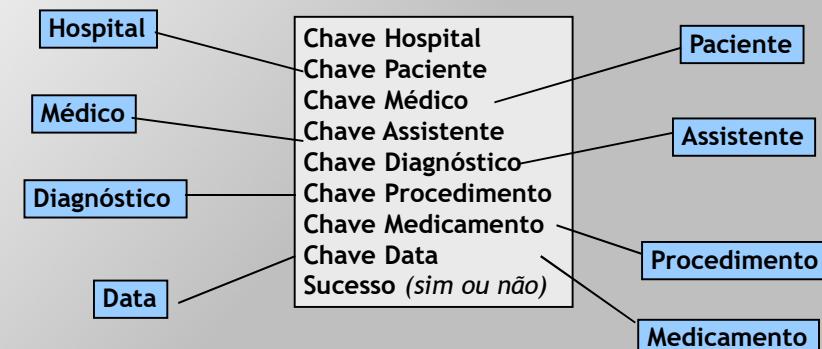
- Ex.: frequência diária a um curso
- Não há fatos
- Ex. contagem da freq.de alunos por prof.



125

Tabela Fato Sem Fatos

❖ Procedimentos Hospital - Paciente

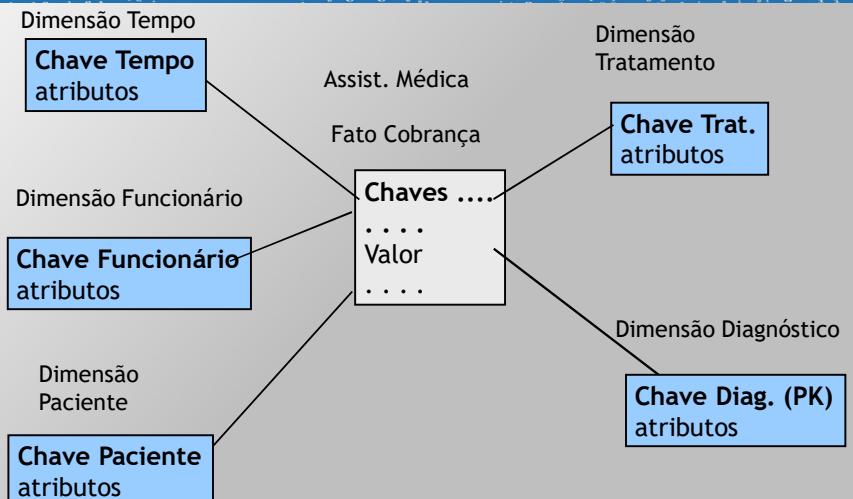


126

Variantes do Modelo Dimensional
Dimensões Multivalor

127

Dimensões Multivalor



128

Dimensões Multivalor

Fato Cobrança

Ch. Paciente	Ch. Diagn.	Valor
A	1	20,00
A	2	50,00
A	3	30,00
B	11	100,00
B	12	30,00
B	3	20,00
B	4	50,00

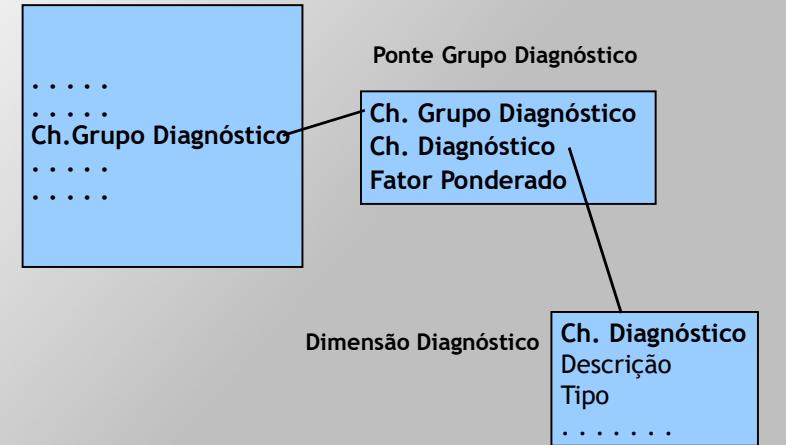
Dim. Diagnóstico

Ch. Diag	Descr.	...
1	X	
2	y	
3	Z	

129

Dimensões Multivalor

Fato Assist. Médica - Cobrança



130

Dimensões Multivalor

Dimensão Tempo

Chave Tempo
atributos

Dimensão
Tratamento

Chave Tratamento
atributos

Assist. Médica

Fato Cobrança

Chaves
Ch. Grupo (FK)
.....
Valor
.....

Dimensão
Funcionário

Chave Funcionário
atributos

Dimensão
Paciente

Chave Paciente
atributos

Dimensão
Diagnóstico

Ch. Grupo (PK)
Ch. Diagn. (FK)
Fator

Chave Diagn. (PK)
Cód. Doença
Descri.
Tipo

Grupo
Diagnóstico

131

Dimensões Multivalor

Fato Cobrança

Ch. Paciente	Ch. Grupo.	Valor
A	I	100,00
B	II	200,00
C	I	500,00
D	II	250,00

Grupo Diagnóstico

Ch. Grupo	Ch. Diagn.	Fator
I	1	0,20
I	2	0,50
I	3	0,30
II	11	0,50
II	12	0,15
II	3	0,10
II	4	0,25

por \$ ou por
probabilidade ou . . .

132

Agregados

Armazenamento de Cubos

133

Agregados - Armazenamento de Cubos

Resumos pré-calculados/ pré-armazenados

Melhorar desempenho

2 técnicas

- Novas tabela fato
- Novos campos (denominados nível)

134

Agregados - Armazenamento de Cubos

❖ Novas tabelas fato/dimensão

➤ Supermercado

- ✓ Agregado: totais de categoria/loja/dia
- ✓ Agregado: totais mensais/produto/loja
- ✓ Agregado: totais de categoria/totais mensais
- ✓

❖ Novos campos denominados nível

- Campos nível nas tabelas dimensão → reg. de fatos agregados na tabela fato original

135

Agregados - Armazenamento de Cubos

❖ Novas tabelas fato/dimensão

Agregado Vendas por Categoria

Chave Tempo
Chave Categoria
Chave Loja
Chave Promoção
valor vendas
unidades vendidas
custo
número clientes

Dimensão Categoria

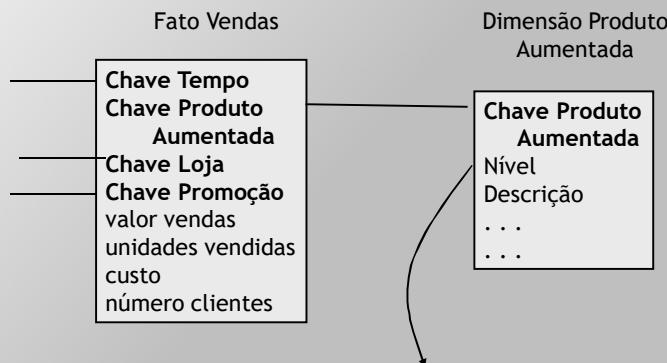
Chave Categoria
Categoria
Deptº

Origem na Dimensão
Produtos
se necessário

136

Agregados - Armazenamento de Cubos

❖ Novos campos denominados Nível



Nível = básico (reg. originais)

Nível = categoria (reg. novos)

137

Agregados - Armazenamento de Cubos

Tabela Produto Original

ID	Nome	Categoria
1	TV	Eletrônico
2	Smart	Eletrônico
3	Freezer	Linha Branca
4	Fogão	Linha Branca

Nova Tabela Categoria

ID	Categoria
A	Eletrônico
B	Linha Branca

138

Agregados - Armazenamento de Cubos

Nova tabela dimensão Produto

Chave Produto Aumentada	Nome	Categoria	Nível	...
1	TV	Eletrônico	básico	
2	Smart	Eletrônico	básico	
A	---	Eletrônico	categoria	
3	Freezer	Linha Branca	básico	
3	Fogão	Linha Branca	básico	
B	---	Linha Branca	categoria	

139

Dispersão/"Explosão" de Agregados

Ex. super
mercado

- 10% dos produtos vendidos em cada dia
- 30.000 produtos com 3.000 grupos de agregados (p.ex. categoria) → aumento de 10% na dimensão produtos
- 25% das categorias de produtos vendidos em cada dia

30.000 produtos
10% vendidos
3.000 reg./dia

+

30.000 produtos
 $\div 10$
3.000 grupos
25% vendidos
750 reg./dia

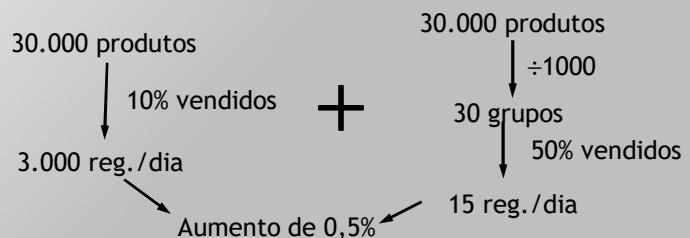
Aumento de 25%

140

Dispersão e Explosão de Agregados

Ex. super
mercado

- Restringir nº de itens de agregação para valor entre 10 e 30
- p.ex. 30.000 produtos → 30 grupos de agregados

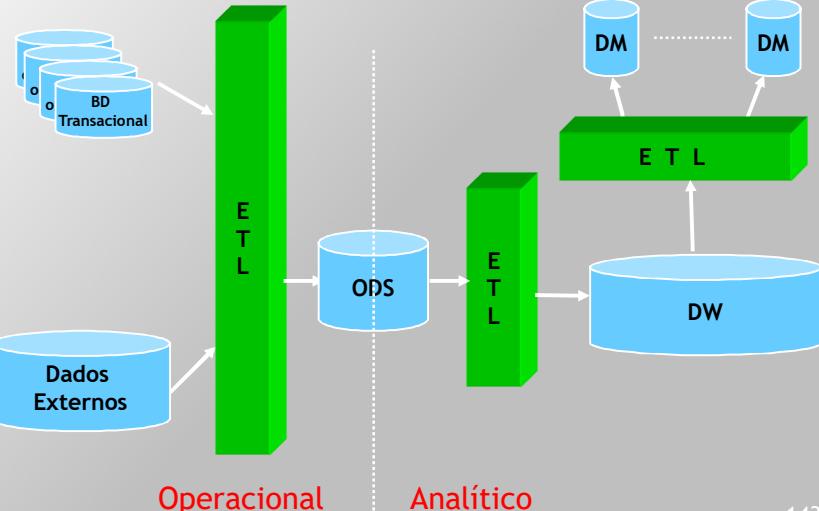


141

ODS - Operational Data Store

142

Estrutura



143

ODS - Características

- Cópias atualizadas de dados operacionais
- Transformação de dados
- Dados detalhados
- Superta atualização
- Qtdade pequena de dados
- Decisões operacionais de curto prazo
- Producir relatórios operacionais

144

Metadados

145

Metadados

- ❖ Dados sobre os dados
- ❖ “Sombra” de dados
- ❖ “DNA” dos dados
- ❖ Em períodos longos de tempo
 - Significado dos dados pode se alterar
 - Controle do histórico de alterações
- ❖ Dados que descrevem
 - Estrutura dos dados
 - Algoritmos de extração/sumarização
 - Mapeamento ambiente operacional → DW (fontes de informação)

146

Metadados

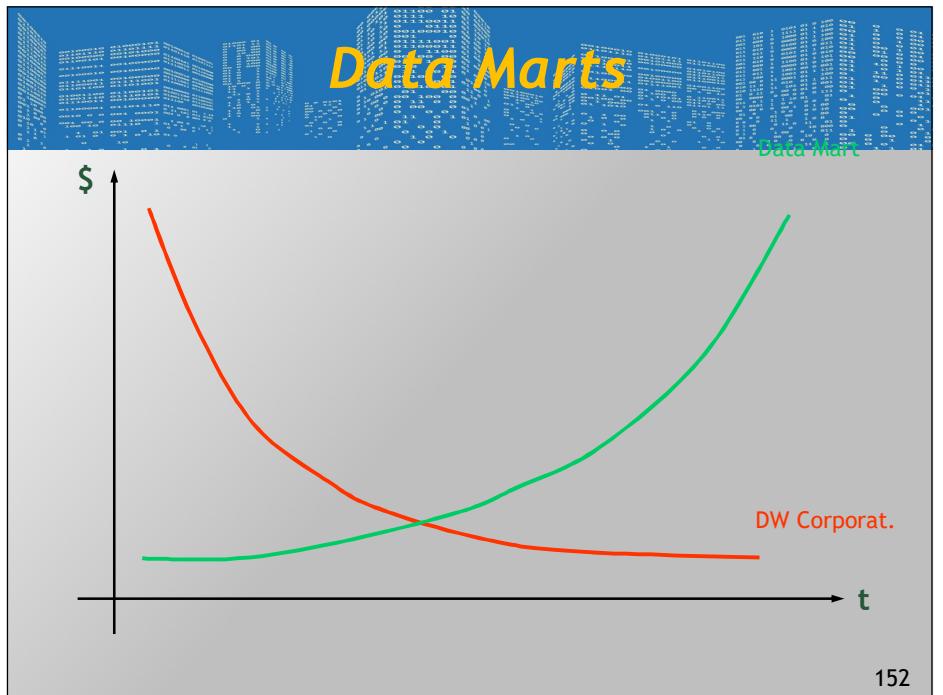
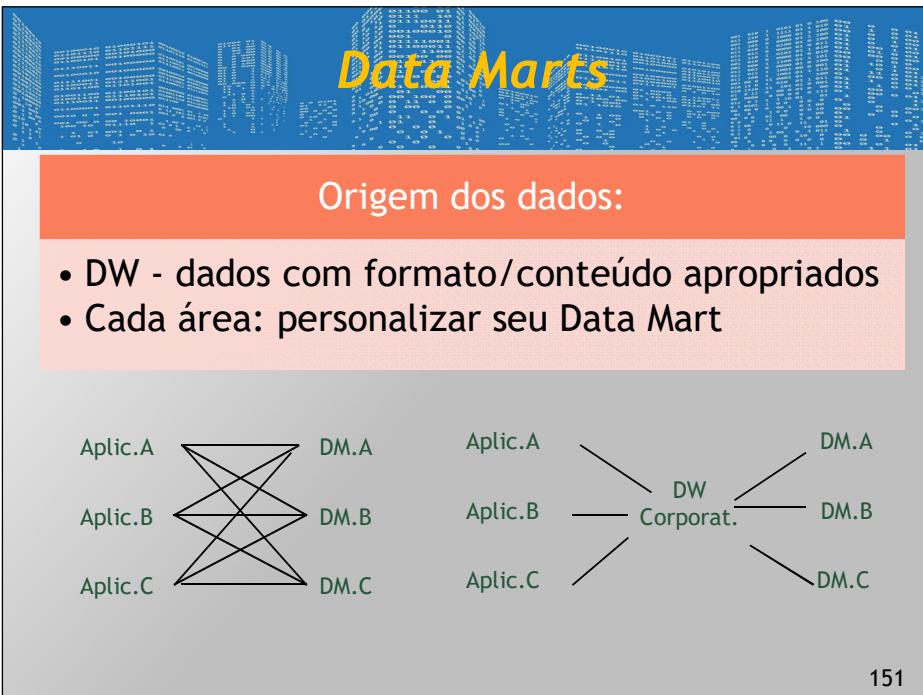
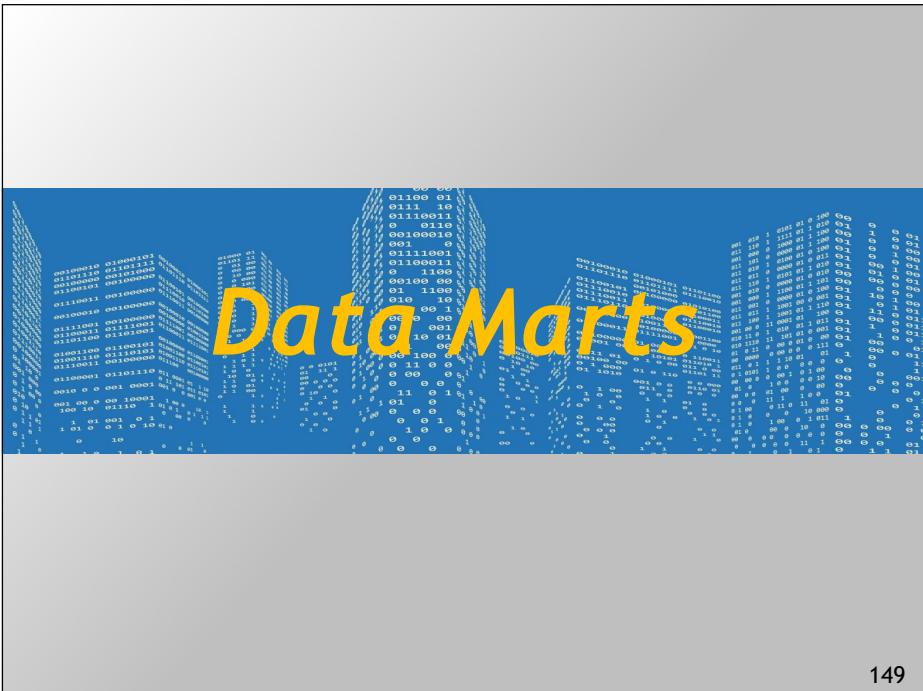
- ❖ Técnicos
 - Estruturas (tabelas, campos, tipos de dados, índices, ...)
 - Origem/Destino dos dados
 - Transformações
- ❖ De Processo
 - Resultado de operações no DW: tempos consumidos no ETL e desempenho de consultas no DW
- ❖ De Negócio
 - Descrição do conteúdo do DW em “alto nível”

147

Metadados

- Descrição dos sistemas fonte do DW
- Responsáveis no caso de problemas
- Estrutura das Tabelas
- Restrições de Integridade

148



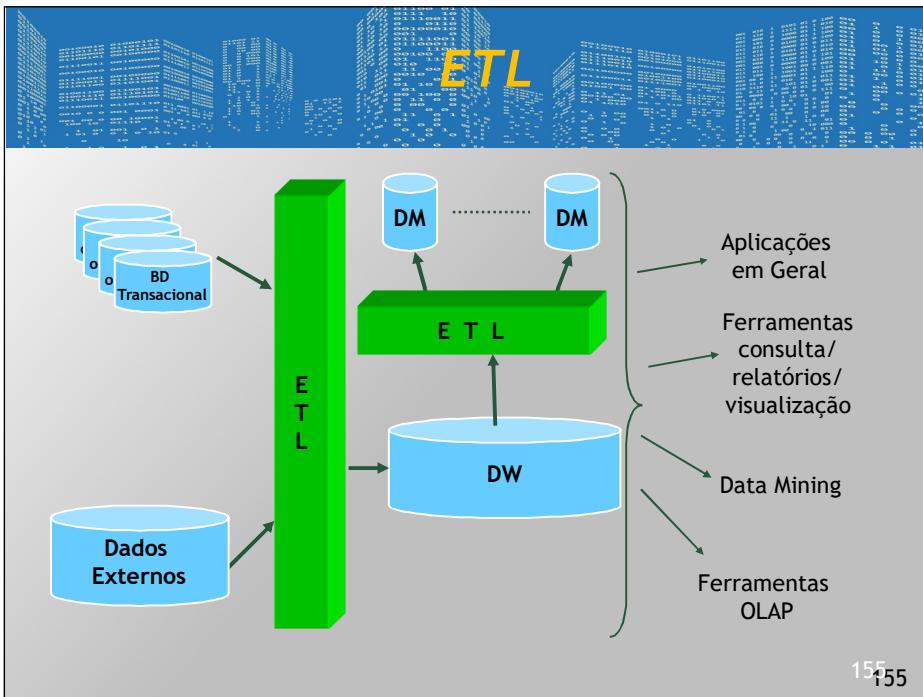
DW x Data Mart

- ❖ Dados da corporação x dados de um grupo ou departamento
- ❖ Assuntos gerais x assuntos específicos
- ❖ Decisões estratégicas x Decisões táticas
- ❖ Modelo dimensional em ambos

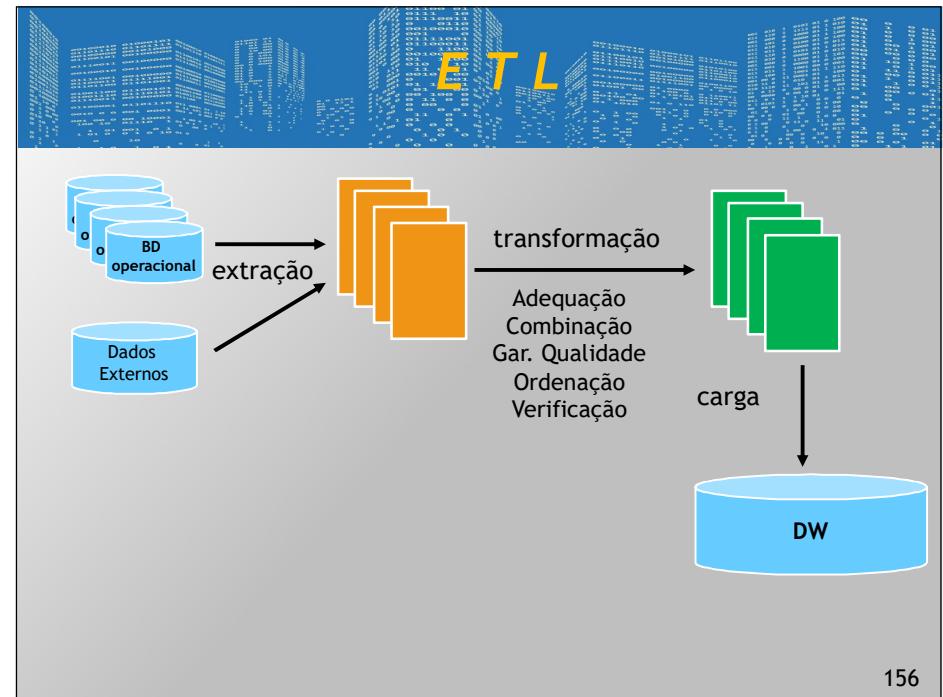
153

Processo ETL

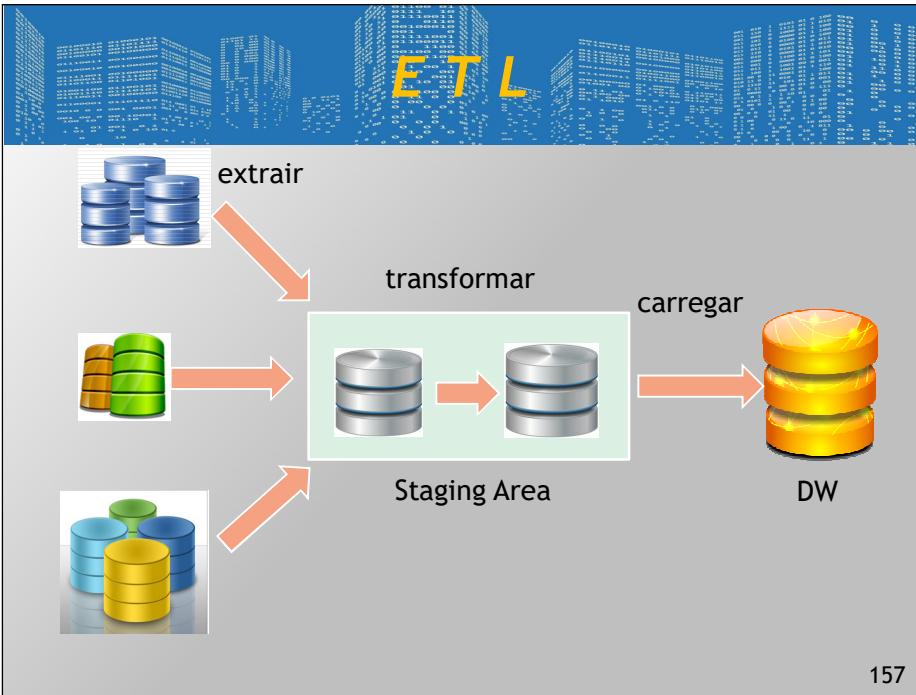
154



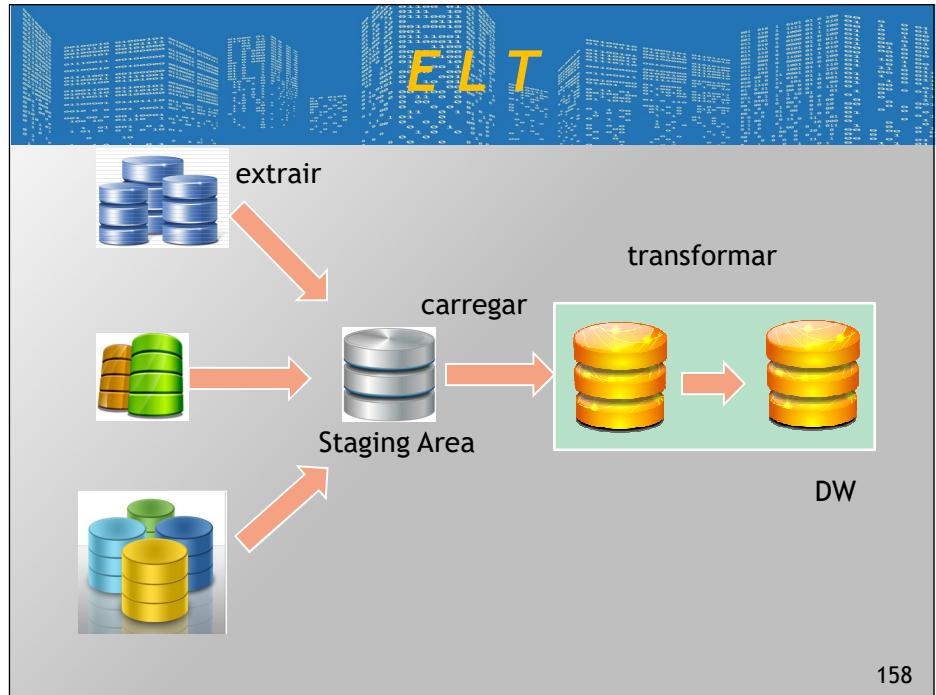
155



156



157



158

NOTE As a rule of thumb, 70 percent of the warehouse development hours are spent on the ETL process. If you are building a new extract system, plan on at least six work months of total ETL development time for each business process dimensional model. Remember that total development time includes understanding source data formats, data profiling, data cleaning, unit testing, system testing, placing into production, and documentation.

The Data Warehouse Lifecycle Toolkit, 2nd edition, Kimball, Ross, Thornthwaite, Mundy, Becker, John Wiley, 2008

159

Note: Unanticipated delays can make the data warehouse project appear to be a failure, but building the ETL process should not be an unanticipated delay. The data warehouse team usually knows that the ETL process consumes the majority of the time to build the data warehouse. The perception of delays can be avoided if the data warehouse sponsors are aware that the deployment of the data warehouse is dependent on the completion of the ETL process. The biggest risk to the timely completion of the ETL system comes from encountering unexpected data-quality problems. This risk can be mitigated with the data-profiling techniques discussed in Chapter 4.

The Data Warehouse ETL Toolkit, Kimball, Caserta, John Wiley, 2004

160

Extrair Dados

- **Data Profiling:** análise técnica dos dados para descrever seu conteúdo, consistência e estrutura
- **Change Data Capture :** identificar os dados que sofreram alterações desde a última carga
- **Extract:** eventuais descompressão e decriptação dos dados

161

Extrair Dados

Data Profiling: análise técnica dos dados para descrever seu conteúdo, consistência e estrutura

Change Data Capture : identificar os dados que sofreram alterações desde a última carga

Extract: eventuais descompressão e decriptação dos dados

Limpar e Processar Dados

Data Cleansing: corrigir dados “defeituosos”, descrição de erros, métricas de QD

Deduplication: eliminar replicações do mesmo dado

Entregar os Dados

Gerenciador de Dimensões de Modificação Lenta

Gerador de Surrogate Keys

162

Aplicações não levam em conta futuras integrações

Dados operacionais

- Mesmos dados com nomes diferentes
- Dados diferentes com mesmo nome

Sinalizar valores “impossíveis”

Fornecer valores padrão

163

Operacional

aplic. A m, f -----> m, f

aplic. B 1, 0 -----> m, f

aplic. C x, y -----> m, f

aplic. D masc., fem. -----> m, f

Codificação

Unidades de Medida Abreviaturas

Descrições distintas Pontuação

Chaves distintas Formas de Codificação

Valores Nulos

Data Warehouse

164



- ❖ Dois tipos de carga: ambiente operacional para o DW



165

- ❖ Aplicação com Marcas de Tempo

- Dados posteriores a um tempo especificado

- ❖ Arquivo Delta

- Contém apenas as alterações ocorridas
- Muito eficiente
- Poucas aplicações geram

- ❖ Arquivo de Log

- Mesmos dados que um arquivo delta
- Finalidade: recuperação de BD operacional

166



Fontes de dados externos

- Jornais/revistas
- Boletins informativos
- Relatórios de consultores

Características de dados externos

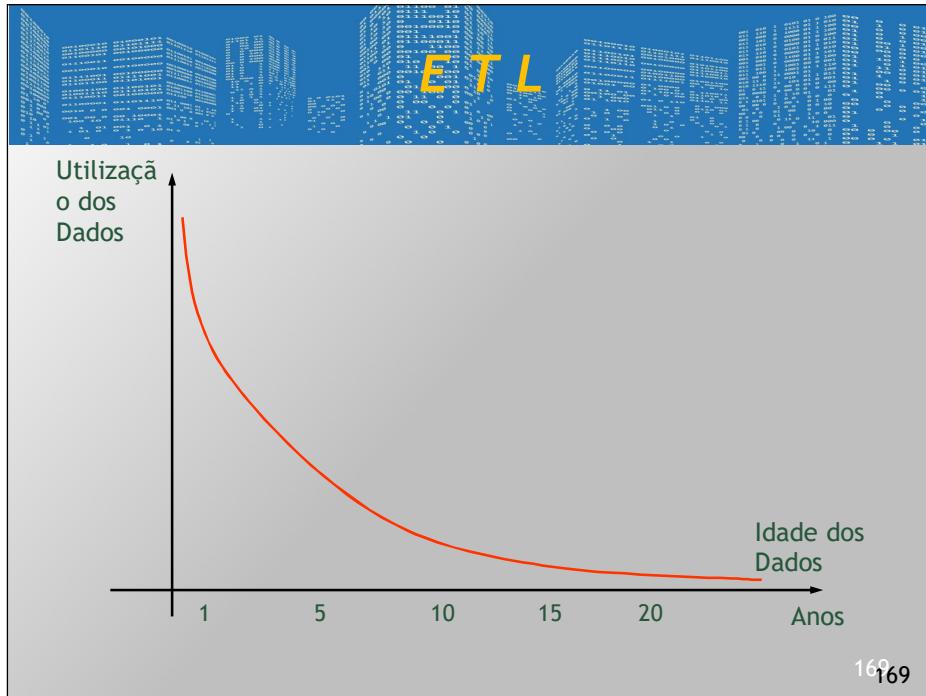
- Periodicidade não é fixa
- Formato variável

167

- ❖ Eliminação ou transferência de dados para outros níveis

- Utilização de resumos
- Transferência para armazenamento de massa
- Eliminação pura

168



169

ETL

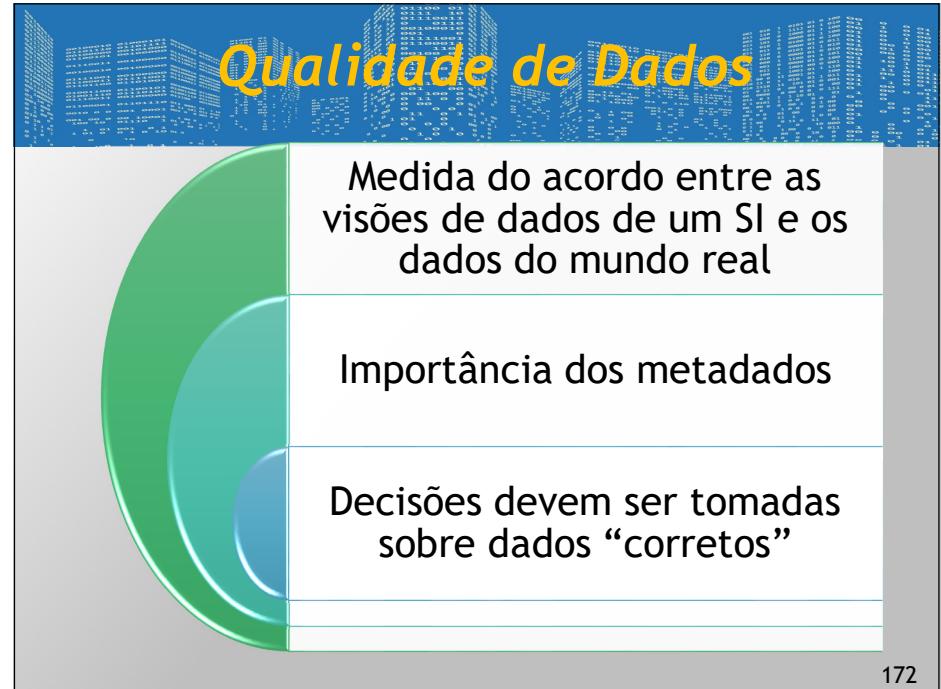
171



170



171



172

Do Data Warehouse para o Ambiente Transacional

173

Do DW para o Amb. Transacional

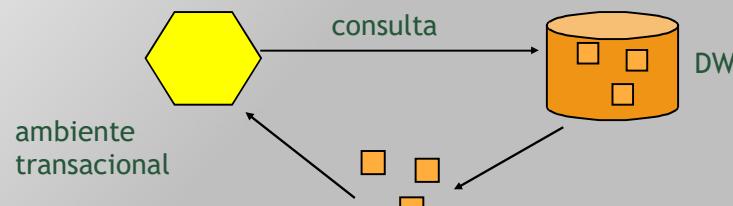
❖ Situação anormal (possível)

- Sequência de condução dos negócios
- Alto desempenho do ambiente operacional
- Obsolescência dos dados

174

Do DW para o Amb. Transacional

- ❖ Acesso direto - muito raro
 - Sem exigências de tempo de resposta

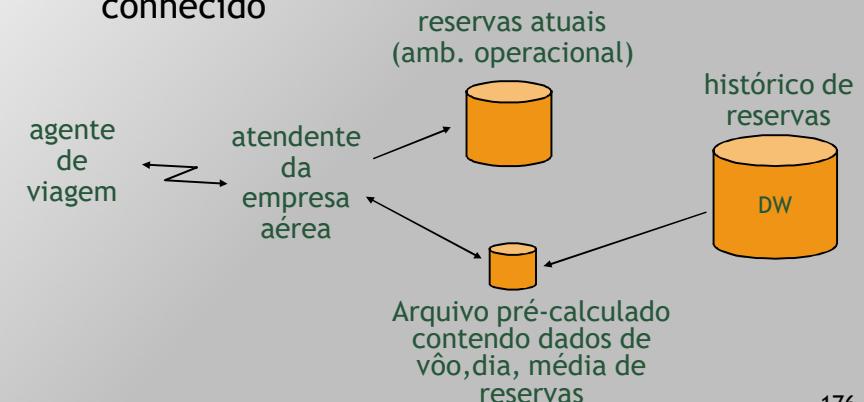


175

Do DW para o Amb. Transacional

❖ Acesso Indireto

- Criação prévia de tabelas resumidas de uso conhecido



176

Monitoramento do Ambiente DW

177

Monitoramento do Ambiente DW

❖ Dois componentes monitorados

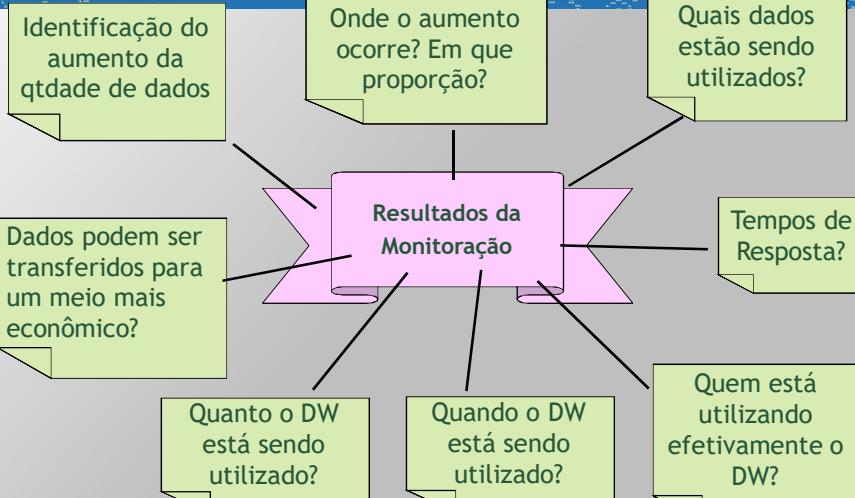
- Quantidade de Dados
- Utilização dos Dados

❖ Crescimento do volume de dados

- Necessidade de recursos adicionais de hardware → maiores custos para DW

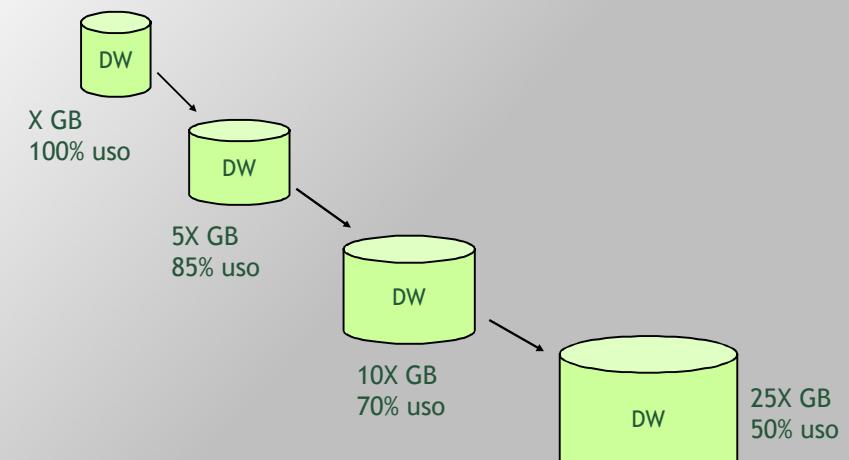
178

Monitoramento do Ambiente DW



179

Monitoramento do Ambiente DW



180

Monitoramento do Ambiente DW

❖ Tempo de Resposta

- Extremamente crítico no amb. operacional
- Não é fator limitante do DW (não pode ser desprezado)



181

Monitoramento da Atividade

Monitoramento

Utilização dos dados:
registros e atributos
acessados

Tempo de resposta

Usuários: quais e
freqüência

183

DWA - Data Warehouse Administrator

❖ Manter organização do DW

❖ Habilidades

- Projetar/construir o DW
- Política - competir pelos recursos necessários
- Gerente de Sistemas - selecionar HW/SW

❖ Atividade de Monitoração do DWA

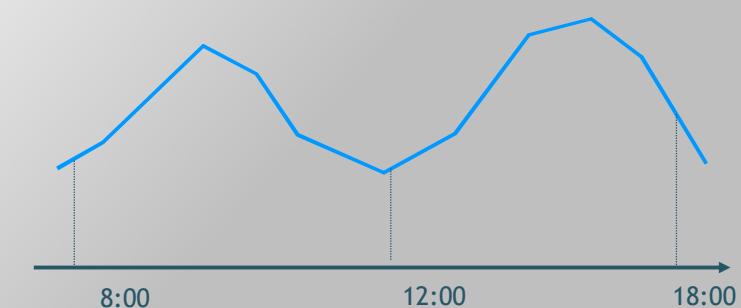
- Determinar tabelas a serem monitoradas
- Determinar freqüência de monitoração
- Carga do software de monitoração no DW

182

Monitoramento da Atividade

❖ Monitor de atividades

- Consultas: tipo e tamanho
- Ocorrência de atividades



184

Monitoramento da Atividade

❖ Granularidade: medir atividades em

- Tabelas - Atributos - Registros

❖ Carga do Monitor

- Processamento extra no DW
- Maior parte do monitoramento em nível pouco detalhado
- Aumento de detalhes nas tabelas mais críticas

185

Monitoramento da Atividade - Chargeback

Usuário final toma conhecimento do custo relativo de cada consulta

Consciente dos recursos consumidos

Algoritmo - parâmetros

- Horário de submissão de consultas
- Tamanho/Prioridade de consultas
- Nº de consultas submetidas
- Período da semana/mês de submissão

186

Monitoramento de Dados

❖ Grandes volumes de dados

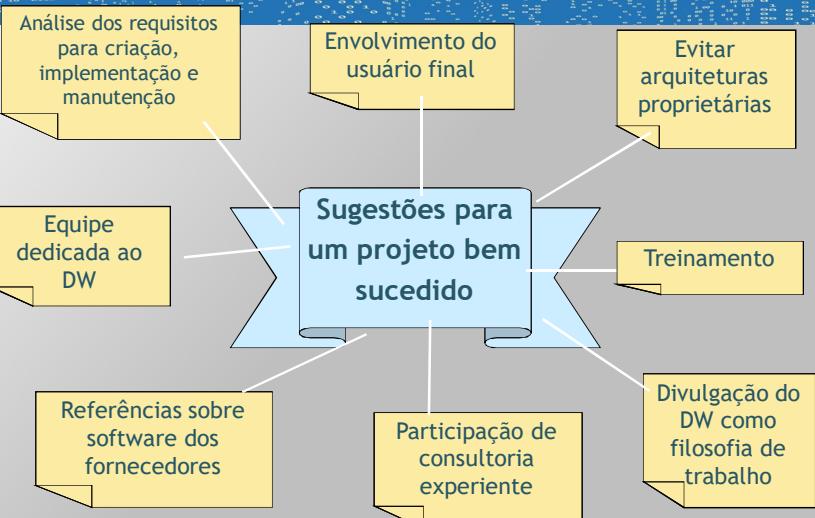
- Monitoração do DW não é feita de uma só vez
- Divisão para execução em horários alternativos
 - ✓ Tempo
 - ✓ Setores
 - ✓ Regiões geográficas
 - ✓ Faixas de venda
 - ✓ ...

187

Projeto do Data Warehouse

188

Implantação



189

Desenv. Iterativo - Justificativa

- Histórico de sucesso das aplicações
- Usuários não expressam com precisão suas necessidades até a primeira iteração:
“Dê-me o que eu digo que quero e, então poderei lhe dizer o que realmente quero”
- *“Dê-me o que eu digo que quero e, então poderei lhe dizer o que realmente preciso”*

190

Desenv. Iterativo - Justificativa

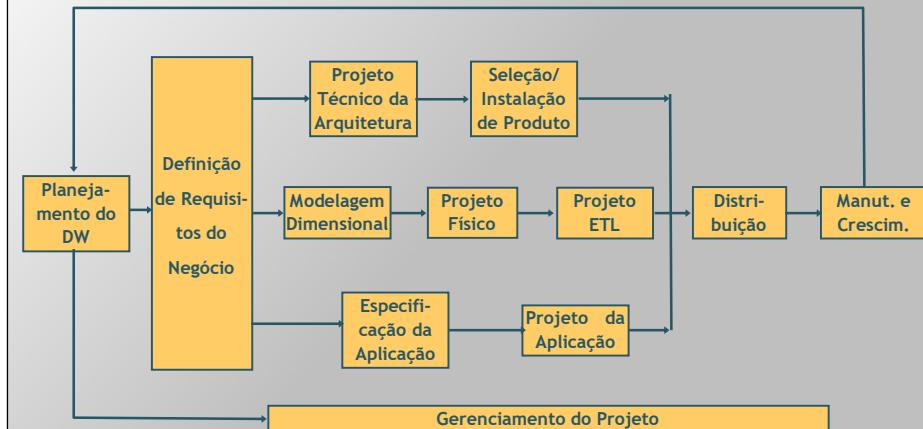
Gerência: espera pela apresentação de resultados concretos

Difícil estabelecer reais benefícios do DW

Necessidade de obtenção rápida de resultados

191

Ciclo de Vida



192

DW - Passos para o Projeto

Especificação do DW

Projeto da Matriz de Barramento

Projeto da Tabela Fato

Definição dos Fatos

Granularidade da Tabela Fato

Definição das Dimensões

Dimensões de Modif. Lenta

Amplitude de Tempo

Intervalo de Extração

DW - Passos para o Projeto

❖ Obter dados necessários

- Documentação dos Ambientes Operacionais
- Entrevistas com Usuários Finais e DBA

❖ Entrevista com DBA

- Informações sobre Ambiente Operacional

➤ Perguntas típicas

- ✓ Como os vários sistemas de produção se relacionam entre si?
- ✓ Quem mantém o arquivo master dos produtos?
- ✓ Obter metadados desses sistemas

194

Modelo Fato Dimensional

195

Modelo Fato Dimensional

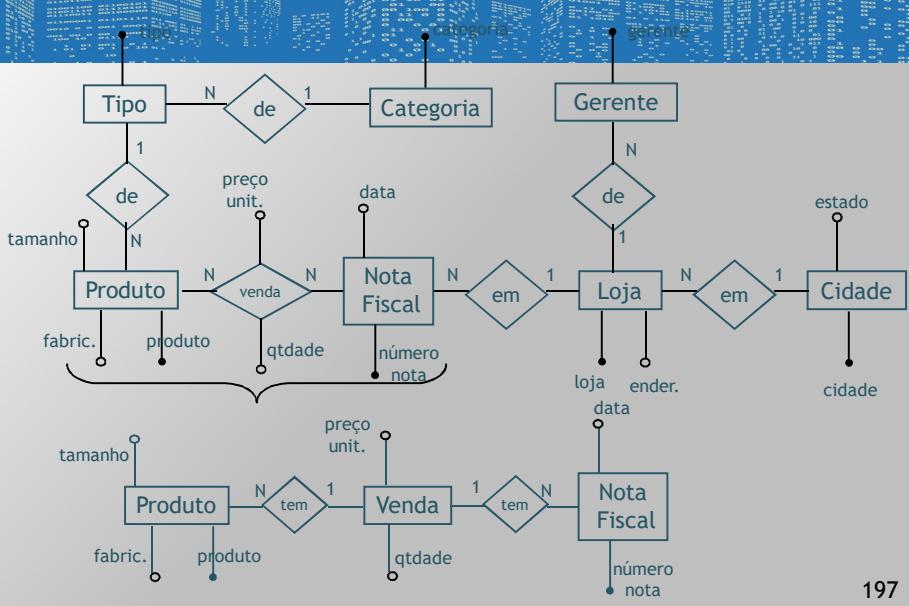
❖ Conceptual Design of Data Warehouse from E/R Schemes

Matteo Golfarelli, Dario Maio, Stefano Rizzi

Proceedings of the Hawaii International Conference on Systems Sciences, January 6-9, 1998

196

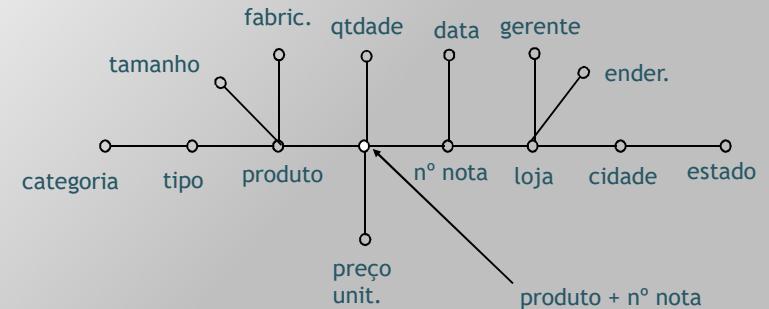
Modelo Fato Dimensional



197

Modelo Fato Dimensional

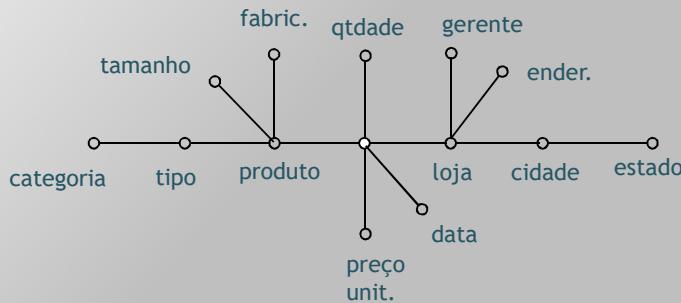
Montar a árvore de atributos



198

Modelo Fato Dimensional

Fazer cortes e enxertos



199

Modelo Fato Dimensional

Hierarquia

Dimensão



200

Análise

201



- ❖ Permite acesso rápido a dados críticos em um único local
- ❖ Acesso a relatórios estruturados e ad-hoc
- ❖ Redução do tempo total para análise
- ❖ Acesso a dados históricos permite analisar diferentes períodos de tempo

202

Desvantagens do DW

- ❖ Não suporta dados não estruturados
- ❖ Difícil implementação de alterações na implementação dos modelos
- ❖ Grande esforço para treinamento e implementação

203



- ✓ Custos de manutenção
- ✓ Propriedade dos dados - segurança
- ✓ Necessidade de transformação dos dados origem (ETL)
- ✓ Subestimar o tempo necessário ao ETL
- ✓ Longa duração para a implantação do DW

204

Especificação do data warehouse

205

Especificação de um DW

- ❖ Inclusa na especificação do Sistema Computacional
- ❖ Modelagem
 - Objetivo do DW
 - Funções do DW
 - Quais os dados de origem
 - Quais as consultas estruturadas
 - E os agregados ?
 - Quais as consultas ad-hoc
 - Projeto da Matriz de Barramento e dos modelos dimensionais

Modelagem

- Tipos de dados, definição de chaves
- Metadados
- ETL - principalmente o T
- Intervalo de extração
- Forma de descarte de dados “vencidos”

206

Especificação de um DW

Modelagem

➤ Definir Área de Aplicação

1. Definir Indicadores (Fatos) a serem analisados; gerais e específicos → Tabelas Fato
2. Definir Fatores (Dimensões) a serem considerados → Tabelas Dimensão
3. Determinar a origem dos dados
4. Projetar a Matriz de Barramento
5. Projetar o Modelo Dimensional macro
6. Transformação dos dados
7. Detalhar consultas de cada Modelo Dimensional

207

Especificação de um DW

Desempenho

- Iniciação
- Consultas
- Carga de dados
- Geração de cubos
- Backup

Hardware

- Servidores
- Discos
- Backup
- No-break

Software

- SGBD
- Ferramentas ETL
- Ferramentas OLAP
- Ferramentas de geração de relatórios/gráficos
- Ferramentas de Data Mining

Equipe DW

- Gerência
- DWA
- Analistas de desenvolvimento e manutenção

208