



INE5643 - Data Warehouses

Aulas 7 a 9 – Modelagem Dimensional

**Prof. Mateus Grellert
Prof. Renato Fileto**

**Departamento de Informática e Estatística (INE)
Universidade Federal de Santa Catarina (UFSC)**

Tópicos

- I. O Modelo de dados dimensional**
 - **Fatos e dimensões**
 - **Esquemas em estrela, floco de neve ou hipercubos**
 - **Medidas de fatos e funções de agregação**
 - **Hierarquias, níveis e membros de dimensões**
 - **Operadores OLAP sobre esquemas dimensionais**
- 2. Projeto de esquemas dimensionais em DWs**
 - Medidas de fatos e funções de agregação
 - Hierarquias, níveis e membros de dimensões
- 3. Projeto físico e de desempenho**
 - Padrões e esquema físico
 - Ajustes para eficiência: indexação, agregações, ...

Posição da modelagem dimensional no ciclo de vida de DW



OLTP vs OLAP

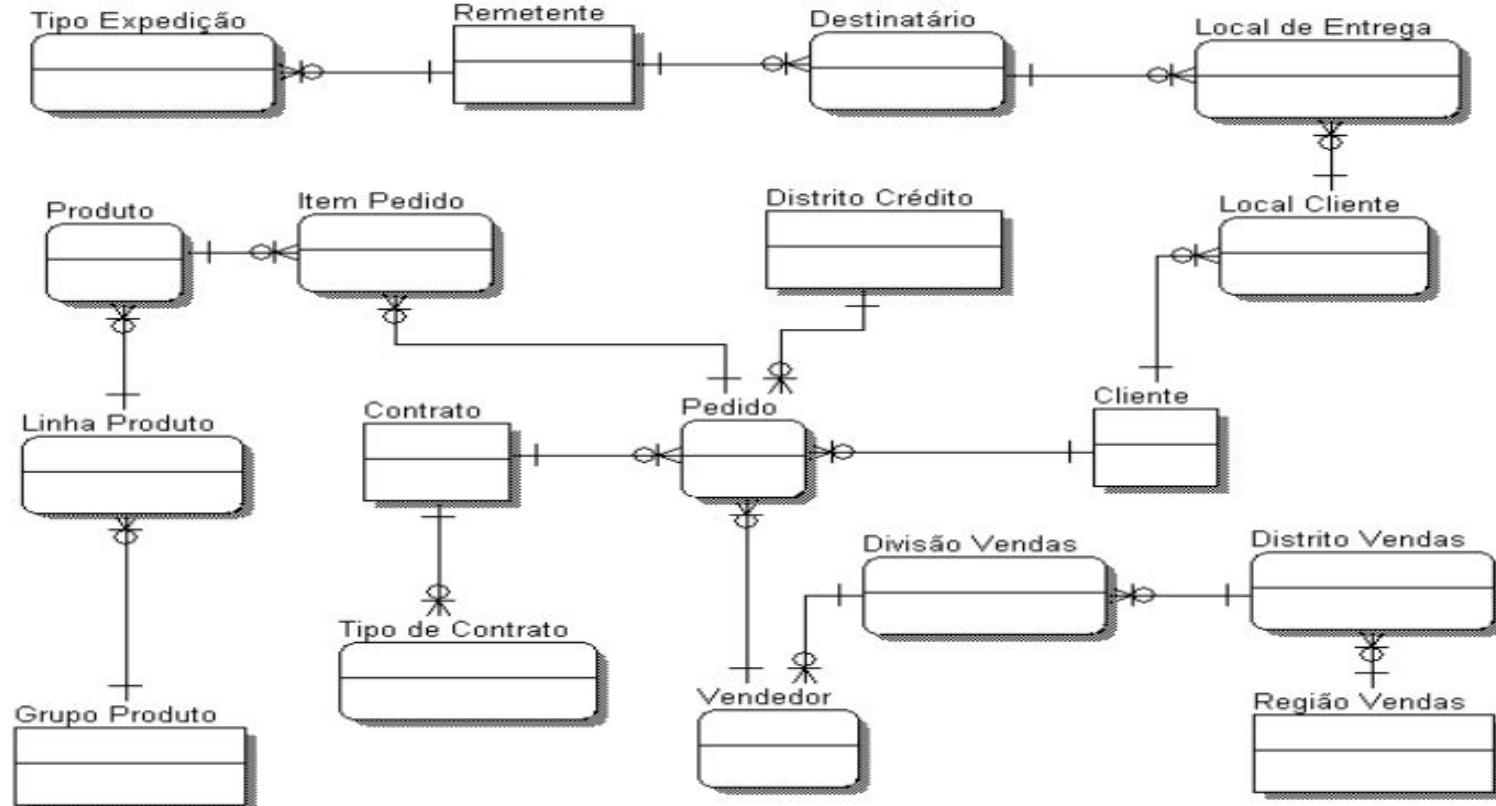
Sistemas Transacionais

- OLTP - Online Transaction Processing
=> Modelagem Entidade Relacionamento

Sistemas Informacionais

- OLAP - Online Analytical Processing
=> Modelagem Dimensional

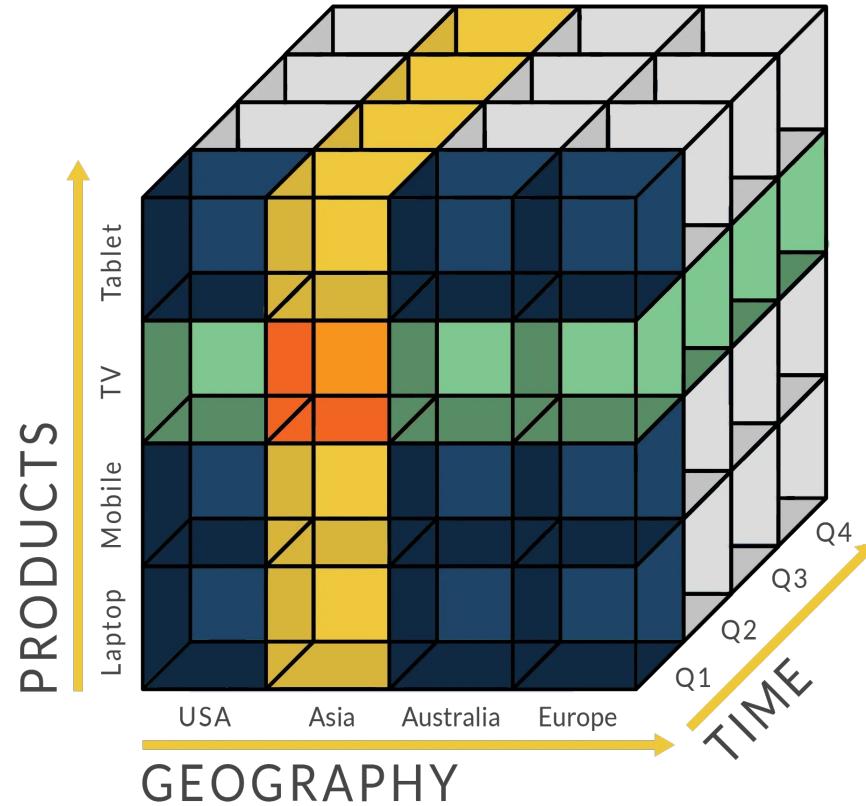
Exemplo de esquema de BD transacional (normalizado)



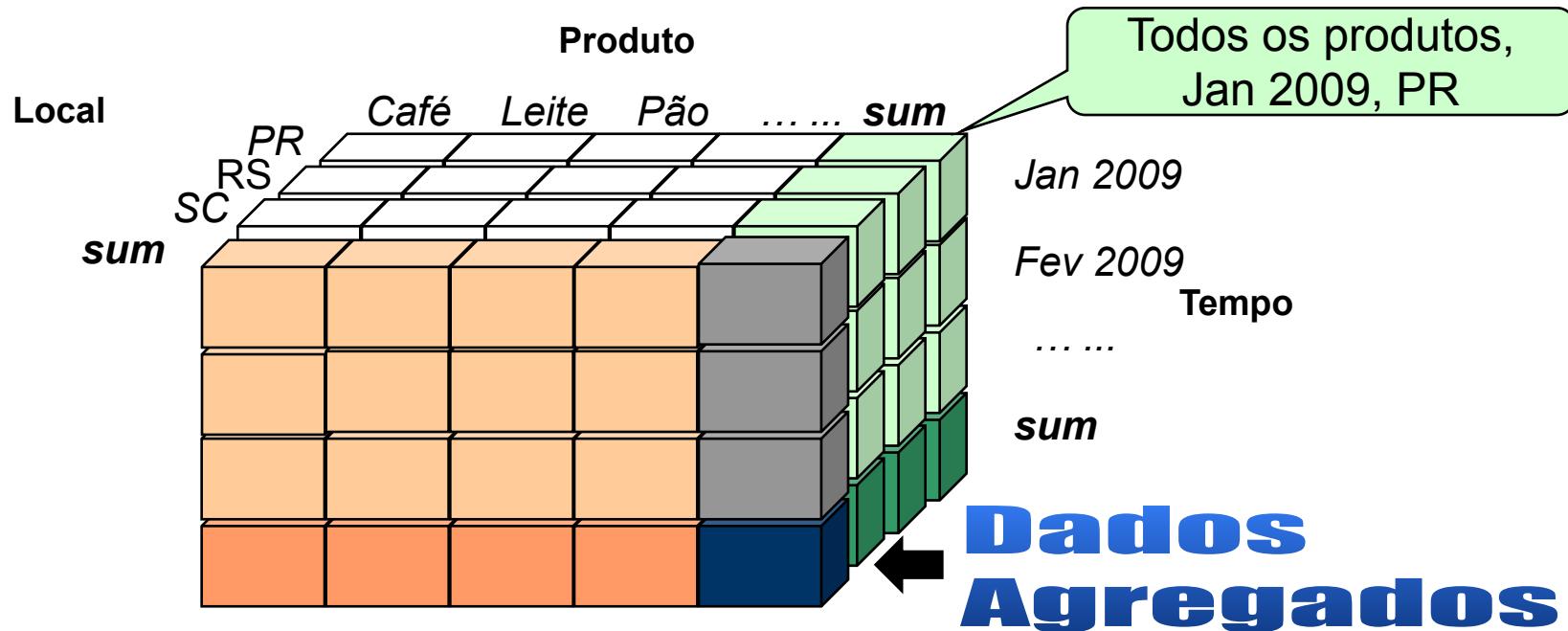
O Modelo de dados dimensional

- Modelo específico para processamento analítico de informação (OLAP)
- Fatos (medidas contextualizadas) segundo dimensões e suas hierarquias, usualmente organizadas em níveis
 - Exemplos de fatos
 - **quantidade** vendida
 - **valor** vendido
 - **número** de habitantes
 - Exemplos de dimensões
 - **Local** com os níveis país, estado e município
 - **Tempo** com os níveis ano, mês e dia
 - **Produto** com os níveis tipo e nome

Dimensional cube



Cubo dimensional vendas



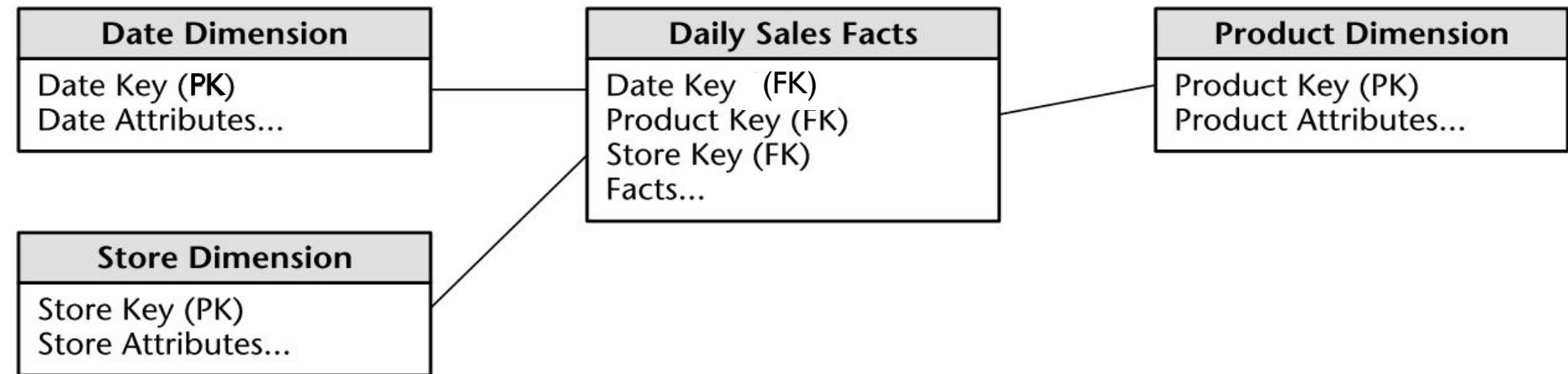
- Células ordinárias (brancas) têm dados no nível mínimo de granularidade para todas as dimensões
- Faces coloridas têm dados agregados (count, sum, max, etc.) nas respectivas dimensões

O Esquema de um DW

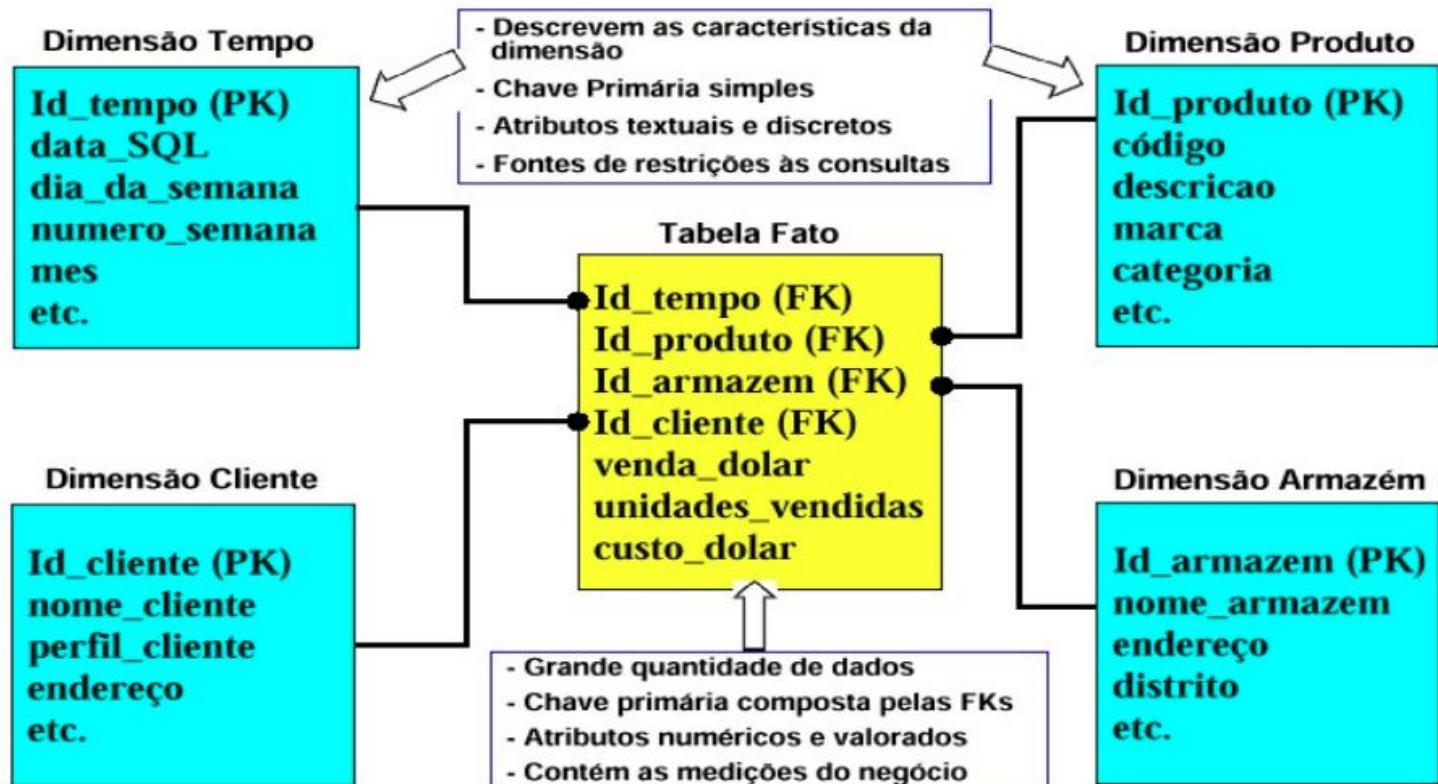
- **Tabela(s) fato** – *Dados quantitativos* – fatos (medidas contextualizadas) coletados originalmente por processos de negócios, redes de sensores ou outras fontes (usualmente dados numéricos, bastante dinâmicos e muitos registros)
- **Dimensões** – *Dados qualitativos* - organizando conceitos (classes) e/ou instâncias para contextualizar, selecionar e agregar fatos, rotulando esses fatos e os resultados de suas agregações (usualmente dados categóricos, mais estáticos e em quantidade bem menor do que os fatos)

Example: Facts and Dimensional Tables

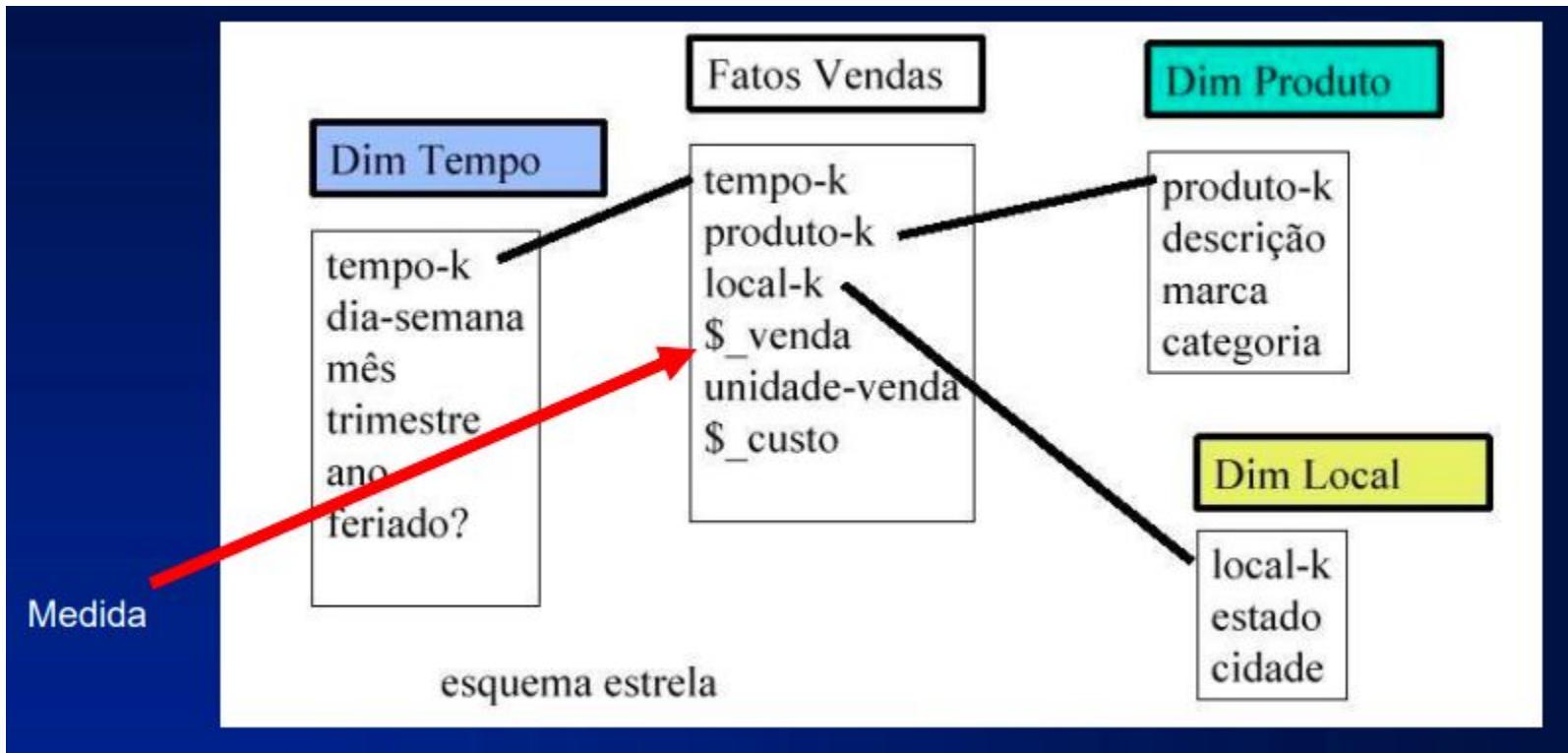
(Kimball 2002)



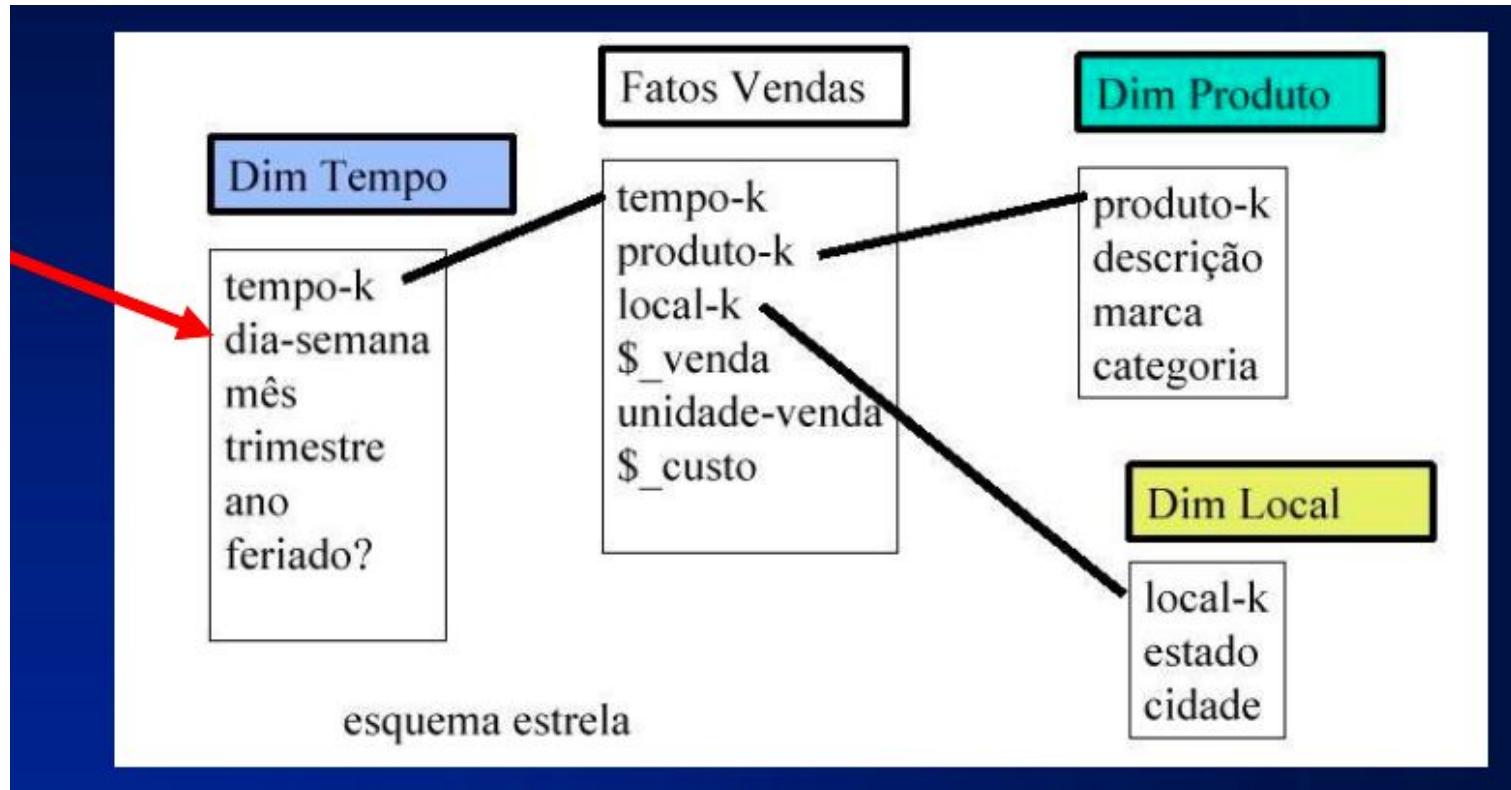
Outro DW sobre vendas



A tabela fato



Uma tabela de dimensão



Vendas de um supermercado



Exemplo de consulta

<u>Marca</u>	<u>Vendas R\$</u>	<u>Unidades Vendidas</u>
Aurora	780	263
Frimesa	1044	509
Perdigão	213	444
Sadia	95	39

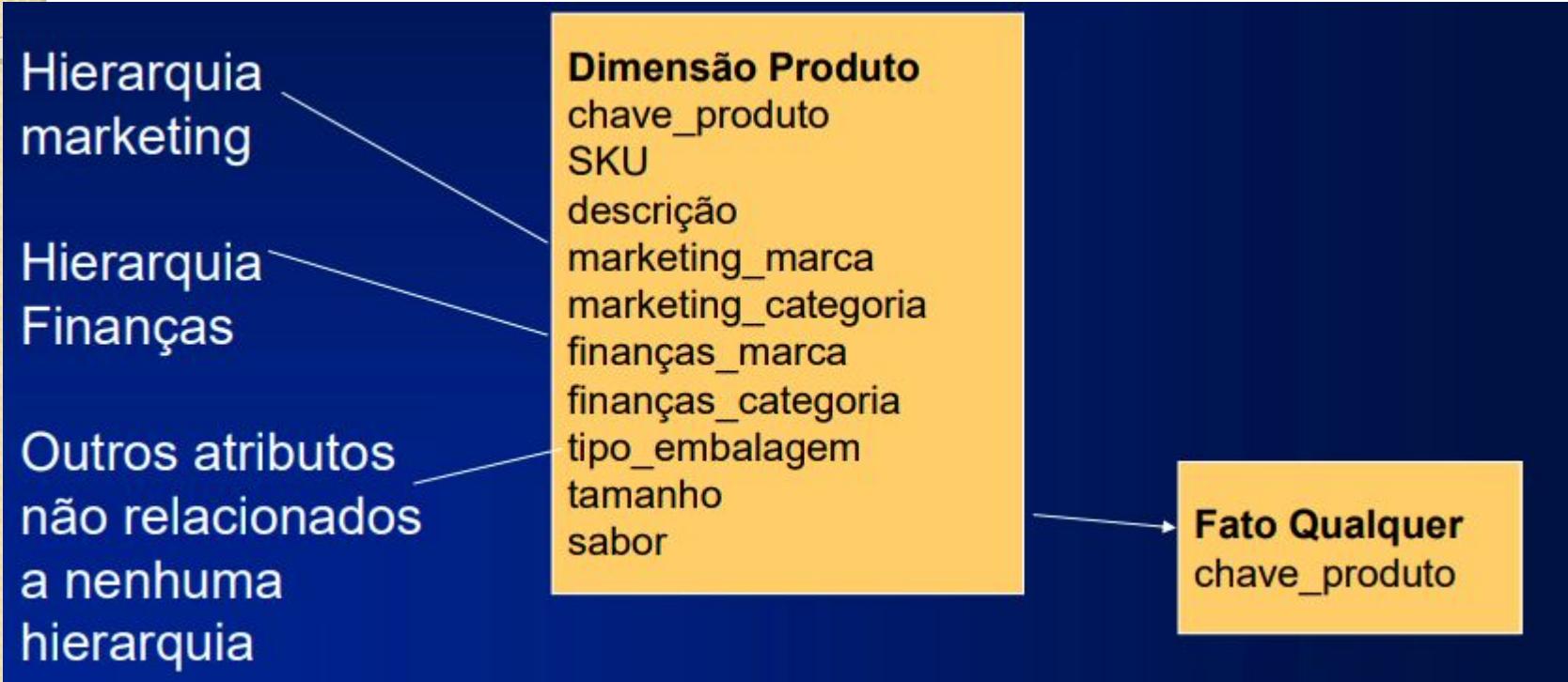
Comandos SQL:

```
Select p.marca, sum(f.total_venda), sum(f.total_unidades) *cabeçalho  
  From fato_vendas f , produto p , tempo t * tabelas  
 Where t.feriado = "Carnaval" * restrição de dimensão  
      and f.chave_produto = p.chave_produto * restrição join  
      and f.chave_tempo = t.chave_tempo * restrição join  
Group by p.marca * instrução group by  
Order by p.marca * instrução order by
```

Uma consulta OLAP sobre um esquema dimensional



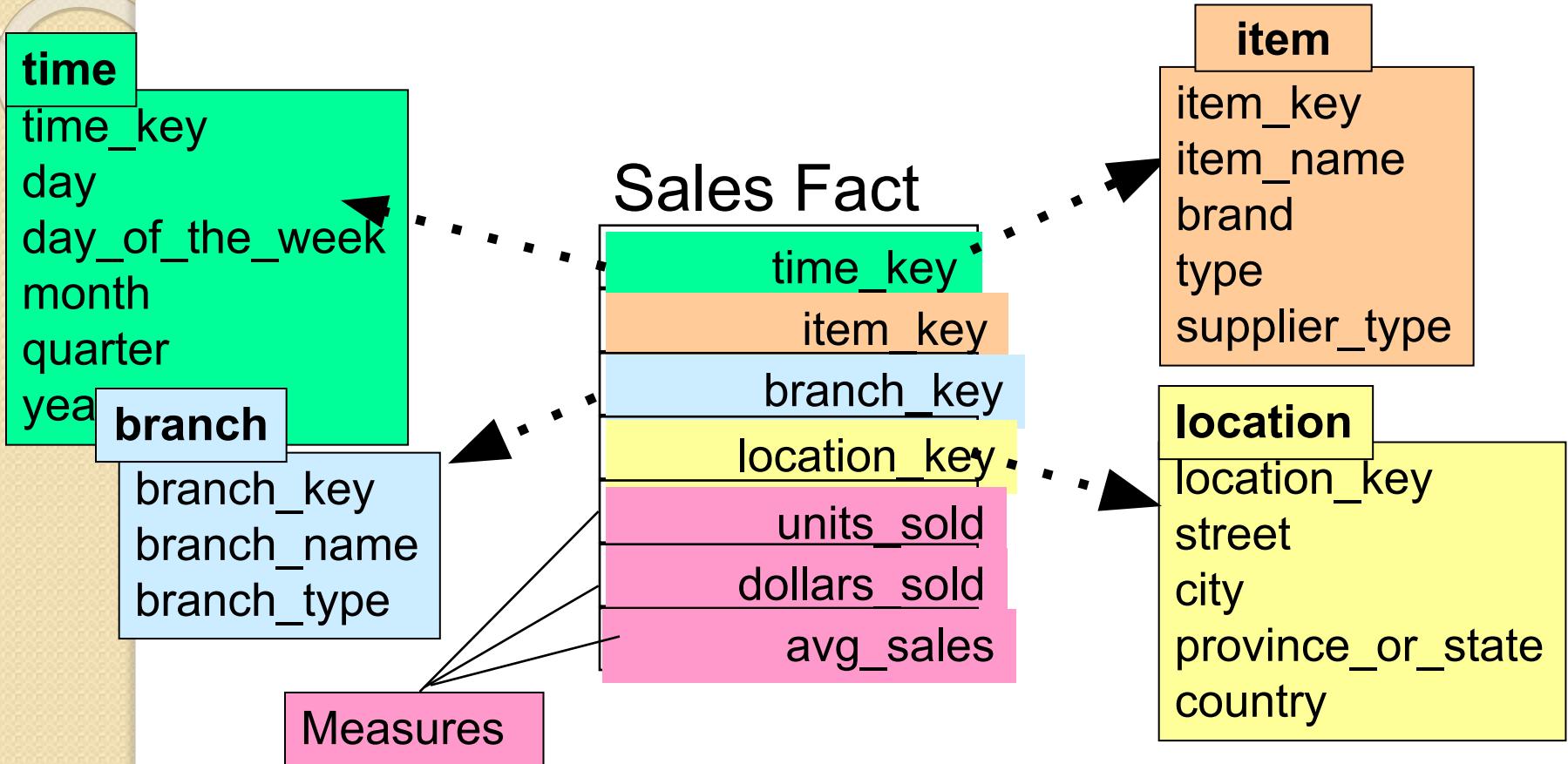
Múltiplas hierarquias em uma dimensão



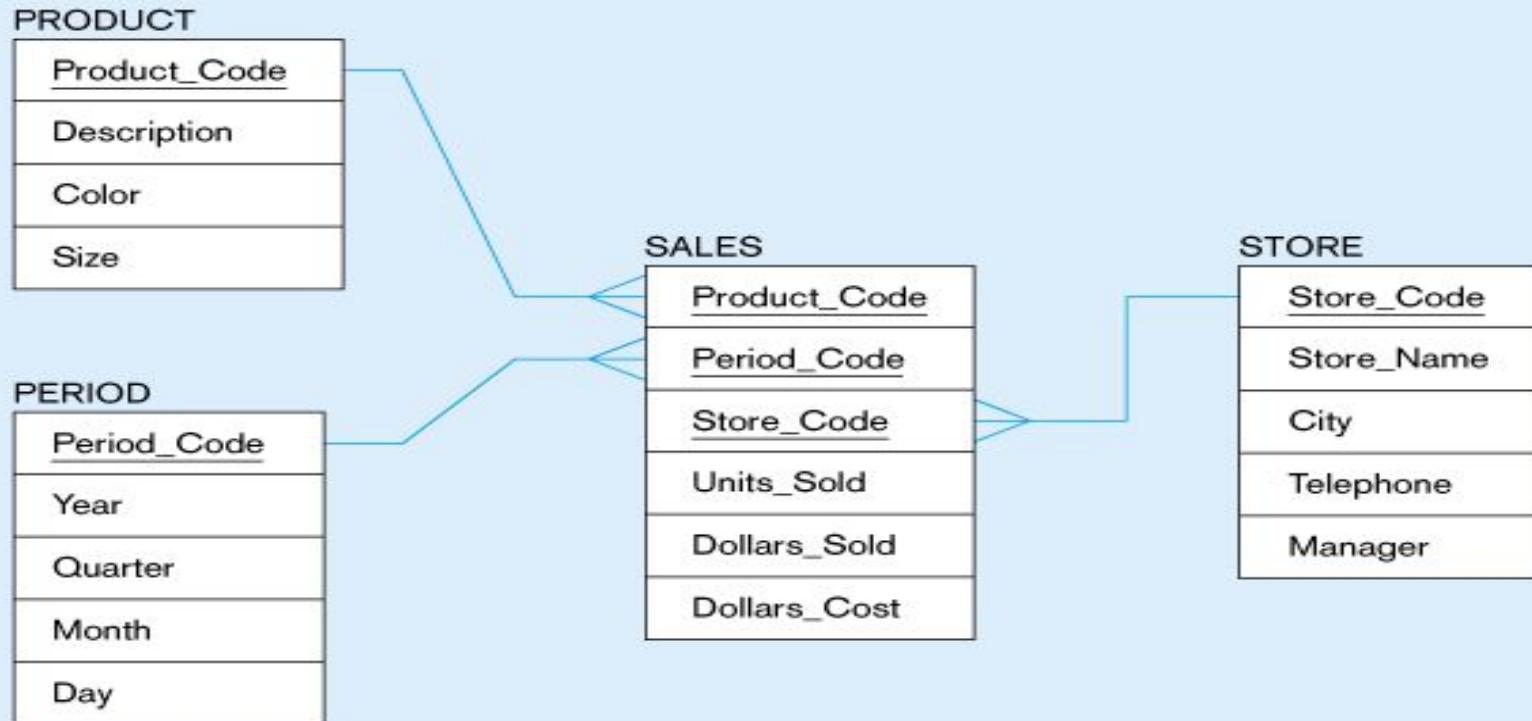
Modelagem de um DW

- **Esquema em estrela (*Star schema*)**
 - uma tabela por dimensão
 - dados não normalizados
 - evita junções entre níveis das dimensões
- **Esquema em floco de neve (*SnowFlake schema*)**
 - permite mais de uma tabela por dimensão
 - dados (parcialmente) normalizados
 - custos com junções
- **Esquema em hipercubo (*Hypercube schema*)**
 - não se ocupa do mapeamento para o modelo relacional

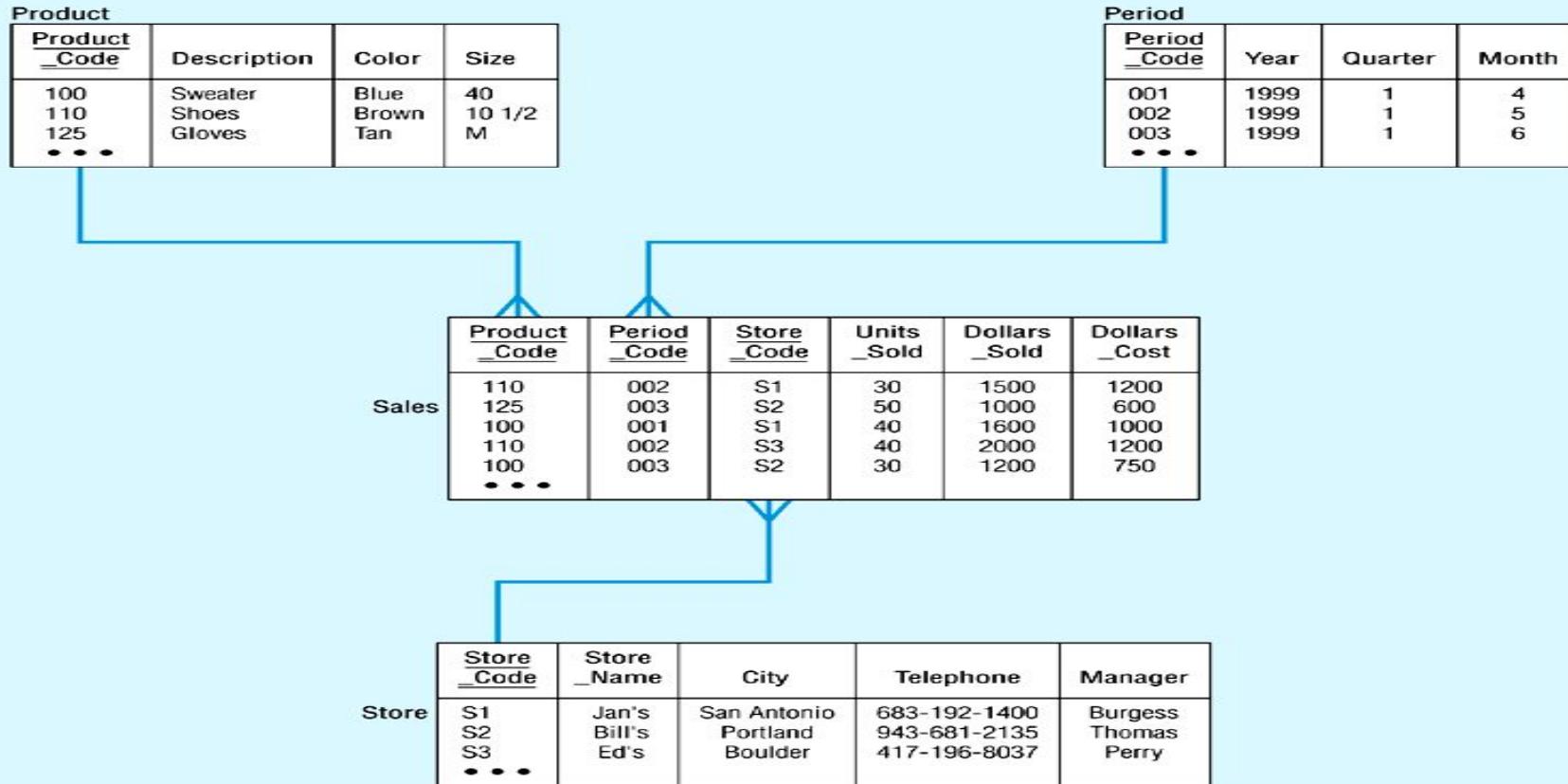
Um esquema em formato estrela



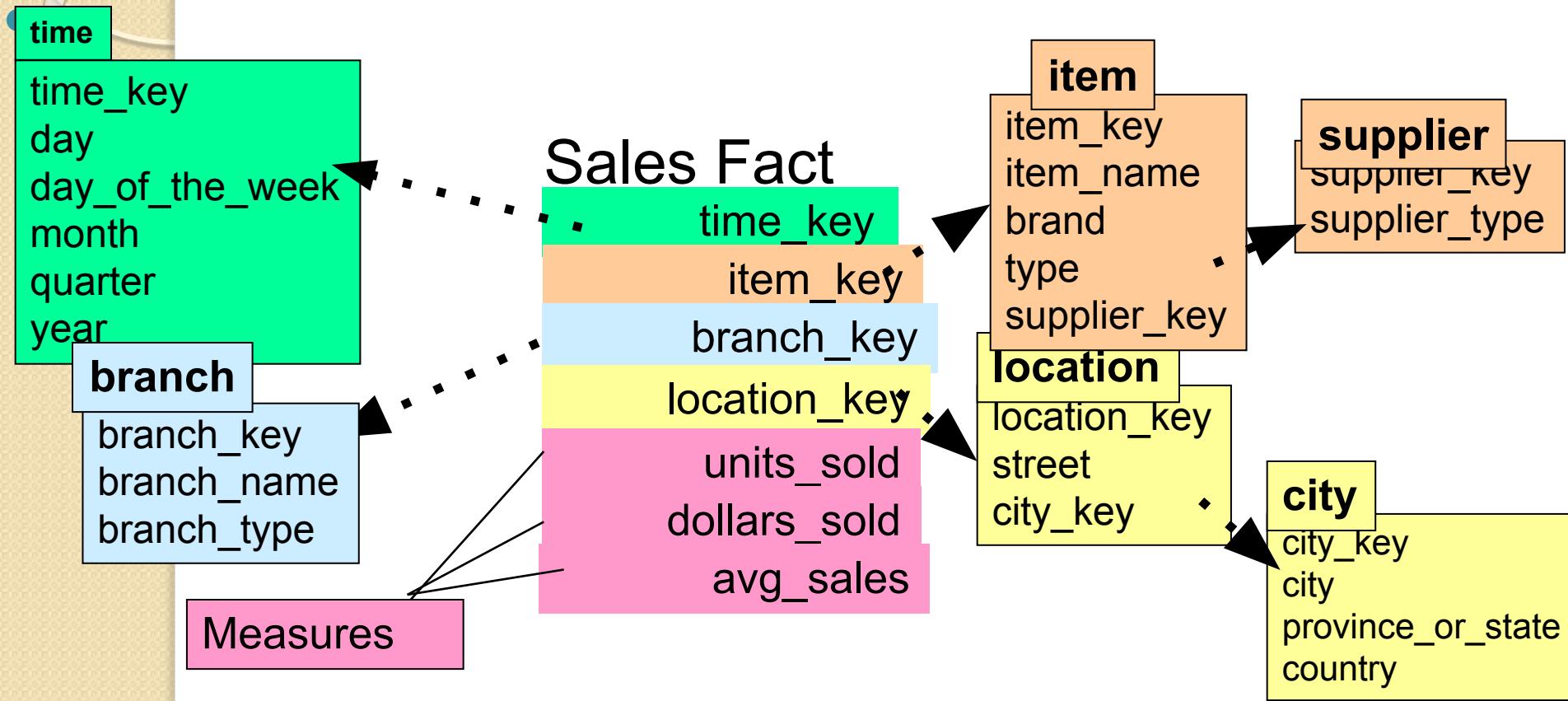
Cardinalidades do esquema estrela



Dados em um esquema estrela (não normalizado)

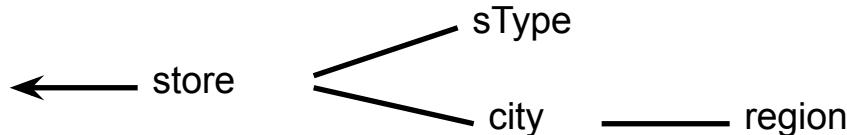


Um esquema em formato floco de neve (normalizado, total ou parcialmente)



Hierarquias de dimensões

(podem ou não ser normalizadas)



store	storeId	cityId	tId	mgr
	s5	sfo	t1	joe
	s7	sfo	t2	fred
	s9	la	t1	nancy

sType	tId	size	location
	t1	small	downtown
	t2	large	suburbs

city	cityId	pop	regId
	sfo	1M	north
	la	5M	south

region	regId	name
	north	cold region
	south	warm region

Uma constelação de Tabelas Fato

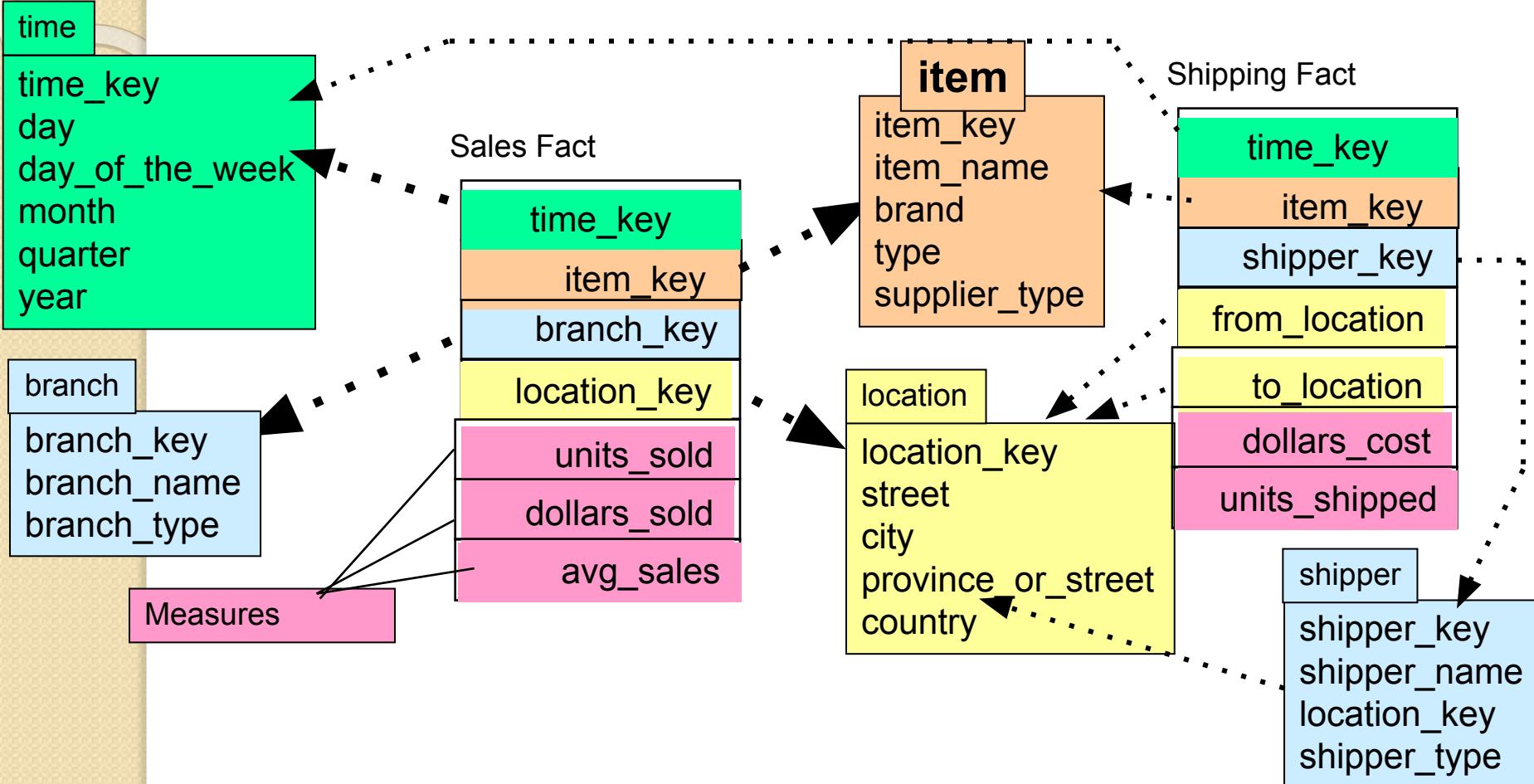


Tabela fato para Visao 2D de Cubo

Tabela Fato

sale	prodId	storeId	amt
	p1	c1	12
	p2	c1	11
	p1	c3	50
	p2	c2	8

Cubo



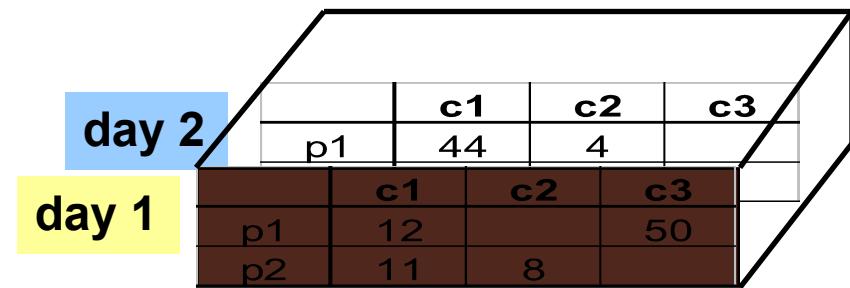
	c1	c2	c3
p1	12		50
p2	11	8	

Tabela fato para cubo 3D

Tabela fato

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

Cubo



Visão da tabela pivot

	Day 1			Day 2		
	c1	c2	c3	c1	c2	c3
p1	12		50	44	4	
p2	11	8				

Agregação de dados

- Quantidade vendida no dia 1

```
SELECT sum(amt)  
FROM SALE  
WHERE date = 1;
```

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4



81

Agregação de dados (II)

- Quantidade vendida por dia

```
SELECT date, sum(amt)  
FROM SALE  
GROUP BY date
```

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4



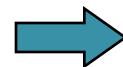
ans	date	sum
	1	81
	2	48

Agregação de dados (III)

- Quantidades vendidas por produto e dia

```
SELECT proId, date, sum(amt)  
FROM SALE  
GROUP BY date, proId
```

sale	proId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4



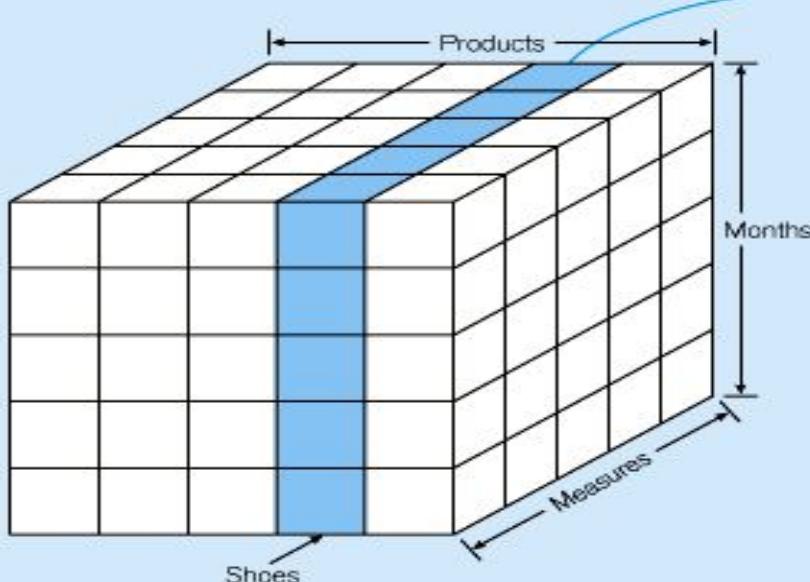
sale	proId	date	amt
	p1	1	62
	p2	1	19
	p1	2	48



Operadores OLAP

- ***Slice***: Projeta valores específicos de uma dimensão (extraita uma fatia do hypercubo)
- ***Dice***: Slices consecutivos (extraita hypercubo menor)
- ***Roll-up (drill-up)***: sumariza dados, subindo na hierarquia de uma dimensão
- ***Drill-down (roll-down)***: reverso de *roll-up*, isto é, detalha os dados, descendo na hierarquia de uma dimensão
- ***Pivot***: muda posição ou orientação da dimensões na projeção bidimensional de dados do hypercubo

Slice



Measure	Units	Revenue	Cost
	January	250	1564
February	200	1275	875
March	350	1800	1275
April	400	1935	1500
May	485	2000	1560

Product: Shoes

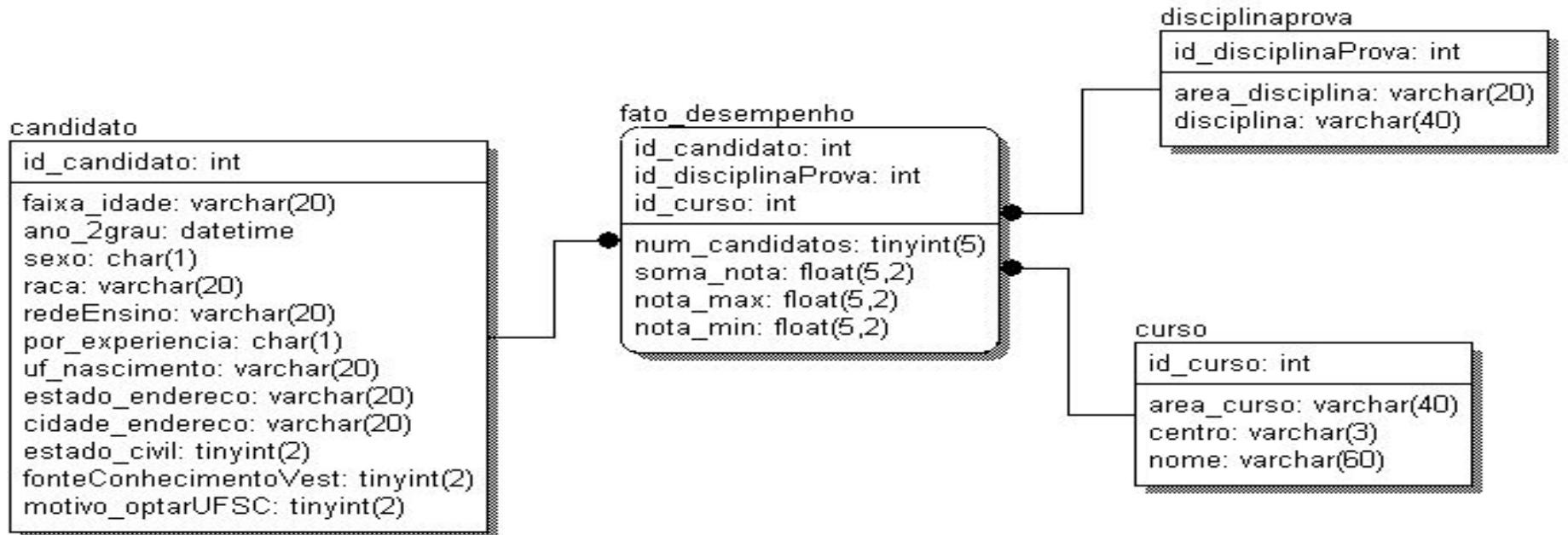
Drill-down

Brand	Package size	Sales
Softowel	2-pack	\$75
Softowel	3-pack	\$100
Softowel	6-pack	\$50

Brand	Package size	Color	Sales
Softowel	2-pack	White	\$30
Softowel	2-pack	Yellow	\$25
Softowel	2-pack	Pink	\$20
Softowel	3-pack	White	\$50
Softowel	3-pack	Green	\$25
Softowel	3-pack	Yellow	\$25
Softowel	6-pack	White	\$30
Softowel	6-pack	Yellow	\$20

Exemplo: DW vestibular UFSC

(Felipe Shigunov, UFSC, 2007)



Visualização no Open

(Felipe Shigunov, UFSC, 2007)



			Measures		
Curso	Candidato	disciplinaprova	Num Candidatos	maxima	Media notas
-All Cursos	+All Candidatos	+All disciplinaprovas	26.538	10,0	3,8
+CCA	+All Candidatos	+All disciplinaprovas	732	9,8	3,6
+CCB	+All Candidatos	+All disciplinaprovas	850	10,0	3,8
+CCE	+All Candidatos	+All disciplinaprovas	2.852	10,0	3,7
+CCJ	+All Candidatos	+All disciplinaprovas	2.051	10,0	4,0
+CCS	+All Candidatos	+All disciplinaprovas	5.516	10,0	4,0
+CDS	+All Candidatos	+All disciplinaprovas	801	9,6	3,3
+CED	+All Candidatos	+All disciplinaprovas	590	9,5	3,2
+CFH	+All Candidatos	+All disciplinaprovas	2.303	10,0	3,6
+CFM	+All Candidatos	+All disciplinaprovas	796	10,0	3,6
+CSE	+All Candidatos	+All disciplinaprovas	3.195	9,8	3,5
+CTC	+All Candidatos	-All disciplinaprovas	6.852	10,0	3,9
		+Ciências Biológicas	1.142	9,8	4,8
		-Ciências Exatas	1.142	10,0	3,7
		FÍSICA	571	9,8	3,9
		MATEMÁTICA	571	10,0	3,6
		+Ciências Sociais	4.568	9,8	3,8

Drill Down

- É usado para solicitar uma visão mais **detalhada** de um conjunto de dados. Pode-se dizer que o usuário "mergulha" nos dados.

A green curved arrow originates from the cell containing '+All Candidatos' in the first table and points to the second table, indicating the transition from a general view to a detailed breakdown.

Drill-Down

Candidato	Measures	
	maxima	num
+All Candidatos	10.00	318,468

Candidato	Measures	
	maxima	num
-All Candidatos	10.00	318,468
+Amapa	6.75	12
+Bahia	9.49	324
+Goias	9.80	1,248
+Mato Grosso	10.00	1,548
+Mato Grosso do Sul	9.80	2,256
+Minas Gerais	9.83	1,704
+Parana	10.00	17,208
+Rio de Janeiro	9.50	552
+Rio Grande do Sul	9.67	10,752
+Santa Catarina	10.00	250,128
+Sao Paulo	10.00	29,484

Roll Up

- Consiste na operação inversa ao Drill-Down, ou seja, apresenta os dados cada vez mais agrupados ou summarizados.

	Measures	
Candidato	• maxima	• num
+All Candidatos	10.00	318,468

Roll Up

	Measures	
Candidato	• maxima	• num
+All Candidatos	10.00	318,468
+Amapá	6.75	12
+Bahia	9.49	324
+Goiás	9.80	1,248
+Mato Grosso	10.00	1,548
+Mato Grosso do Sul	9.80	2,256
+Minas Gerais	9.83	1,704
+Paraná	10.00	17,208
+Rio de Janeiro	9.50	552
+Rio Grande do Sul	9.67	10,752
+Santa Catarina	10.00	250,128
+São Paulo	10.00	29,484

Pivoting

- Serve para adicionar ou rearranjar as dimensões das tabelas

		Measures
Curso	Candidato	• num
+All cursos	+Alagoas	84
	+Amapa	12
	+Bahia	324
	+Goias	1,248
	+Maranhao	12
	+Parana	17,208
	+Sao Paulo	29,484

Pivot

		Measures
Candidato	Curso	• num
+Alagoas	+All cursos	84
+Amapa	+All cursos	12
+Bahia	+All cursos	324
+Goias	+All cursos	1,248
+Maranhao	+All cursos	12
+Parana	+All cursos	17,208
+Sao Paulo	+All cursos	29,484



Slice and Dice

- Para fixar uma informação de dimensão ou reduzir as dimensões de apresentação dos dados

Curso	Candidato	Measures
+All cursos	-All Candidatos	318,468
	+Bahia	324
	+Mato Grosso	1,548
	+Minas Gerais	1,704
	+Parana	17,208
	+Rio de Janeiro	552
	+Santa Catarina	250,128
	+Sao Paulo	29,484

Candidato	Measures
-All Candidatos	318,468
+Parana	17,208
+Santa Catarina	250,128
+Sao Paulo	29,484



**Slice and
Dice**

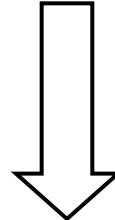


Tópicos

- I. O Modelo de dados dimensional**
 - Fatos e dimensões
 - Esquemas em estrela, floco de neve ou hipercubos
 - Medidas de fatos e funções de agregação
 - Hierarquias, níveis e membros de dimensões
 - Operadores OLAP sobre esquemas dimensionais
- 2. Projeto de esquemas dimensionais em DWs**
 - **Medidas de fatos e funções de agregação**
 - **Hierarquias, níveis e membros de dimensões**
- 3. Projeto físico e de desempenho**
 - Padrões e esquema físico
 - Ajustes para eficiência: indexação, agregações, ...

Projeto DWs

- A **modelagem é crítica** para o sucesso de um DW e merece atenção.
- Empreendimentos que não considerem as **diferenças entre a modelagem de bancos de dados transacionais e DWs**, incluindo questões técnicas e administrativas, podem facilmente fracassar.



Necessidade de critérios, conhecimento e experiência para projeto adequado de data warehouses.



Fases do desenvolvimento de DWs

1. Planejamento
2. Levantamento das necessidades e fontes de dados
3. Integração de dados
4. **Modelagem dimensional**
5. Projeto físico do banco de dados
6. Projeto das transformações de dados (ETC)
7. Desenvolvimento de aplicações
8. Validação e teste
9. Treinamento
10. Implantação

Projeto de um DW

- **Tema**
- **Escopo**
- **Fontes de dados**
- Transformações dos dados
- **Significado dos dados (metadados)**
- **Medidas**
- **Dimensões**
- Análises desejadas

Conselhos para projeto e desenvolvimento de DWs

- **Determine um escopo pequeno**
- Escolha um departamento
- Defina com clareza os objetivos
- Utilize os recursos tecnológicos disponíveis
- **Não proponha um projeto coorporativo**
- **Conceba um projeto escalável**

Formas de Desenvolvimento de DWs

- **Top-down:** Projeto e implementação do DW completo definindo o esquema integrado, fontes de dados e Datamarts.
- **Bottom-up:** Projeto e implementação de pequenas DWs ou DMs que vão se integrando aos poucos.
- **Combinada:** Mistura desenvolvimento de DWs usando várias fontes de dados, diversos DMs e integração incremental.



Modelagem Dimensional

- É crítica para o sucesso de uma DW
- É diferente da modelagem de dados convencional
 - Forma como o usuário visualiza e manipula os dados (organização em hipercubo)
 - A implementação pode ser em SGBDs específicos para DW ou relacionais (verificar a forma como são realizadas junções e outras operações)
 - Diagramas em estrela e floco de neve são utilizados para bancos de dados em hipercubos sobre o modelo relacional
 - Normalização pode ser dispensada (de maneira consciente e controlada) por questões de eficiência

Passos da Modelagem Dimensional

Definir requisitos: a área de negócios (prioridades, mercado, custos e benefícios)

1. Definir os processos dentro da área de negócios
 2. Determinar a granularidade desejada (e viável)
 3. Definir a(s) tabela(s) fato
 4. Descrever as dimensões
 5. Definir as métricas para as medidas dos fatos
 6. Escolher um *DataMart* (definido por uma tabela fato e as dimensões associadas), para começar, projetando e desenvolvendo um por vez
 - Alguns DMs compartilham dimensões e métricas (que podem precisar de ajustes)
- 

Passos da Modelagem Dimensional

Iº Passo: Decidir qual(is) processo(s) do negócio devemos modelar, por meio da combinação do conhecimento do negócio com o conhecimento dos dados que estão disponíveis;

2º Passo: Definir o grão do processo do negócio. O grão é o nível fundamental atômico de dados que representará o processo na tabela de fatos.



Passos da Modelagem Dimensional

3º Passo: Escolher as dimensões que serão aplicadas a cada registro da tabela de fatos.

4º Passo: Escolher os fatos mensuráveis que irão popular cada registro da tabela de fatos.

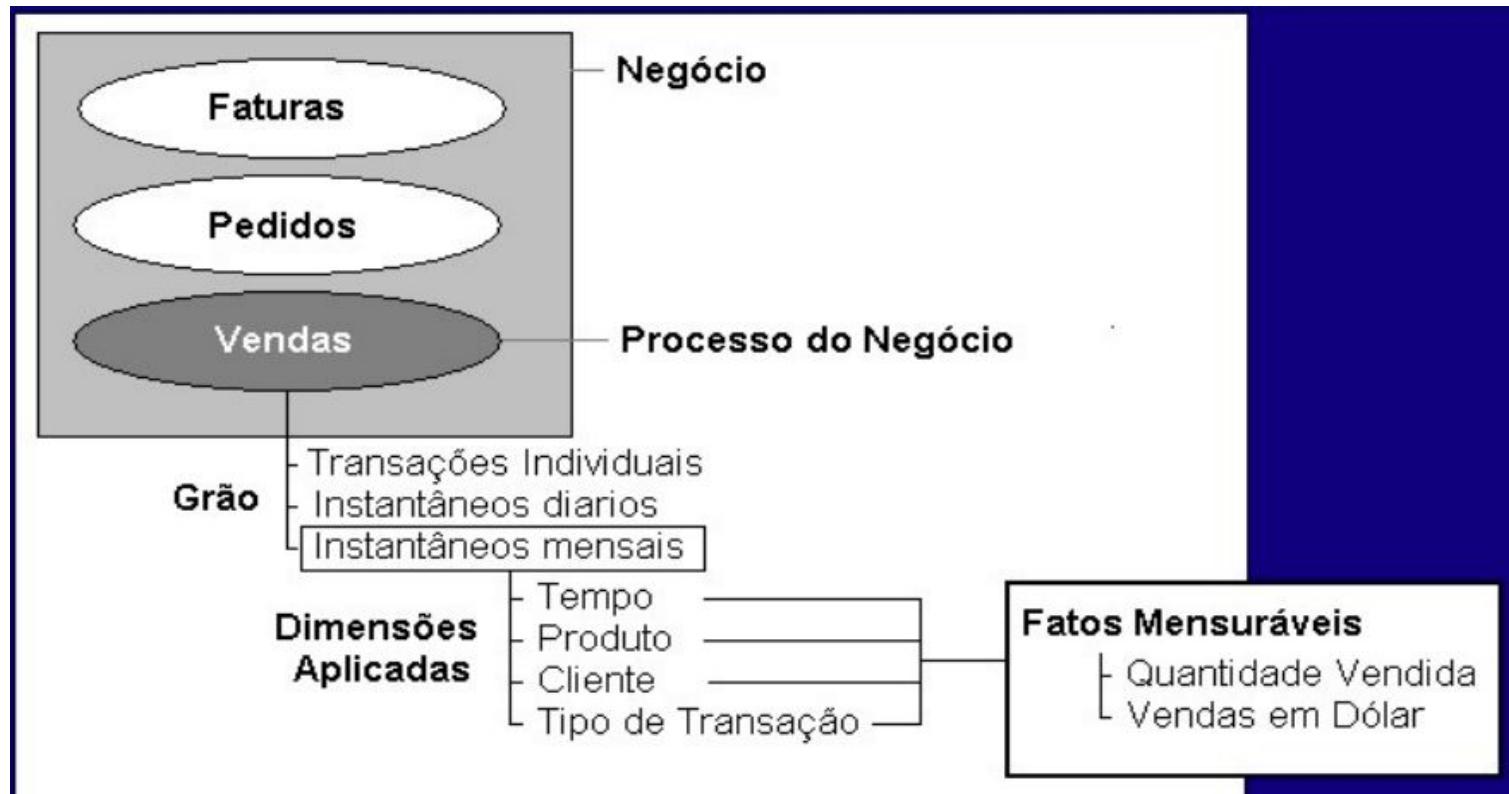


Passos da Modelagem Dimensional

3º Passo: Escolher as dimensões que serão aplicadas a cada registro da tabela de fatos.

4º Passo: Escolher os fatos mensuráveis que irão popular cada registro da tabela de fatos.

Metodologia e critérios



Tradeoff

- **Granularidade Alta:**
 - Economia de espaço em disco;
 - Redução na capacidade de atender consultas.

- **Granularidade Baixa:**
 - Grande quantidade de espaço em disco;
 - Aumento na capacidade de responder qualquer questão.

Métricas de fatos

- **Aditivas:** faz sentido somar
 - quantidade total
 - valor total
- **Semi-Aditivas:** faz sentido somar em certas dimensões
 - quantidade vendida no tempo/espaço
 - quantidade de chuva só no tempo
- **Não Aditivas:** não faz sentido somar
 - $\text{margem_lucro} = \text{preço_venda} / \text{preço_custo}$
- **Não Numéricas:** podem ser unidas, combinadas e contadas
 - pontos espaço-temporais de uma trajetória
 - construções linguísticas de textos

Funções de agregação

- **Decomponíveis (*decomposable*):** permitem calcular agregações de suas agregações
 - SUM, MAX, MIN, COUNT (agrega-se com SUM)
- **Decomponíveis em medidas auxiliares:** podem ser calculadas usando medidas decomponíveis
 - AVERAGE (SUM / COUNT)
 - RANGE ([MIN(min), MAX(max)])
- **Não decomponíveis:** Precisam ser recalculadas por completo para cada contexto, considerando o conjunto inteiro de dados base (não agregados) dentro do contexto
 - DISTINCT COUNT, MEDIAN, MODE



Dimensões

- Referem-se a facetas de contextualização dos fatos
 - O que? (What)
 - Quando? (When)
 - Onde? (Where)
 - Por quê? (Why)
 - Por quem ou qual agente? (Who)
 -



Definindo dimensões

- Na escolha do grão da tabela de fatos algumas dimensões primárias surgem naturalmente;
- Verificar quais dimensões podem ser relacionadas ao grão sem gerar valores duplicados;
- Verificar se a todas as dimensões atendem ao detalhe quantificado na tabela de fatos.

Hierarquia de uma dimensão

- Árvore ou grafo acíclico direcionado (DAG)
 - **Níveis** (usualmente conceitos/classes)
 - **Membros** (muitas vezes objetos/instâncias)
 - **Relações** anti-simétricas entre níveis e entre os respectivos membros, determinando ordens parciais entre níveis e entre os respectivos membros

Dados meteorológicos/climatológicos

Territorial Divisions

- country
- region
- state
- county
- location

Metrics and Measures

- Max Temp
- Min Temp
- Avg Temp
- Total Rainfall

Time

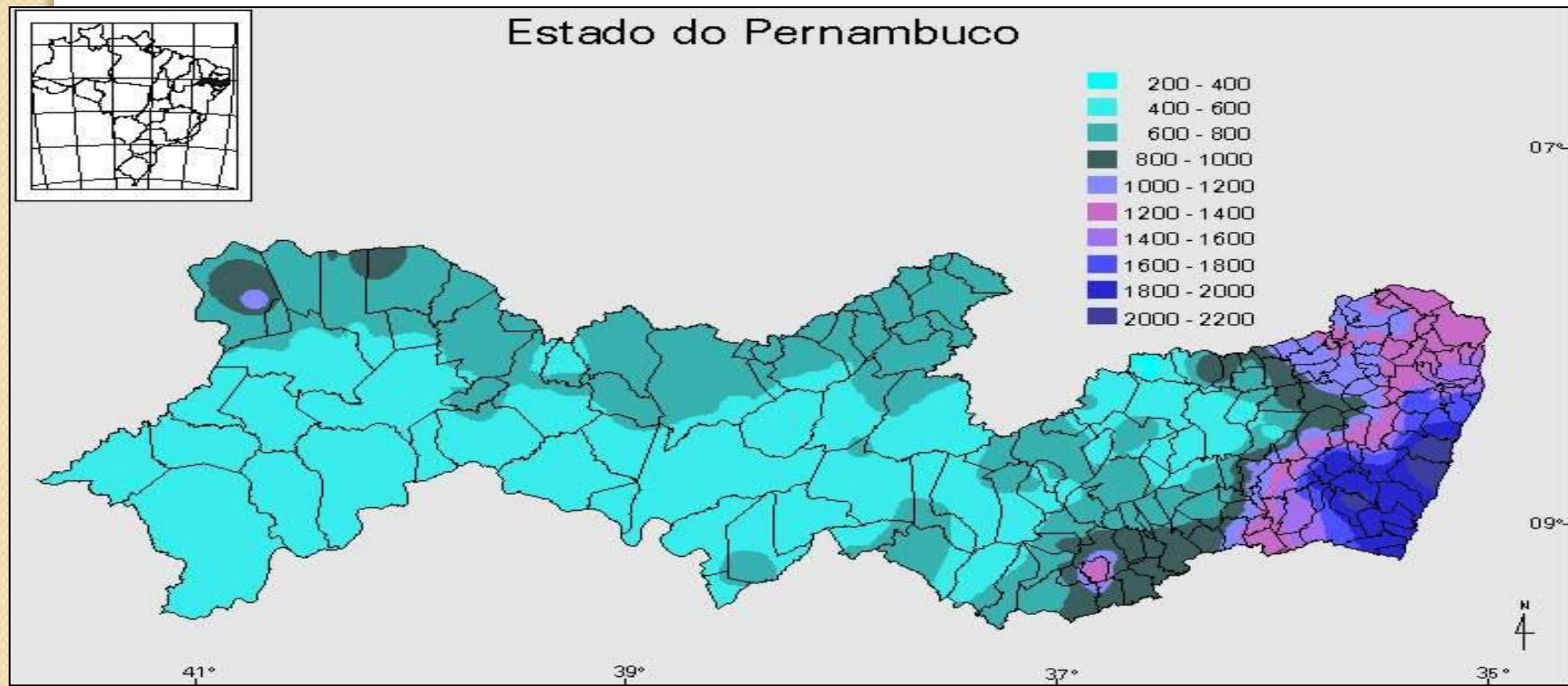
- year
- month
- week
- day

Organizations

- consortium
- institution
- department

Distribuição usual das chuvas em PE

(Cepagri/Unicamp, 2002)



Produção agrícola

Territorial Divisions

- country
- region
- state
- county

Crop Production

- planted area
- expec_production
- monetary_value

Time

- year
- quarter
- month

Products

- class
- family
- crop

Produção de Frutas no Brasil

(Carlos A.A. Meira, Embrapa, 2003)

Product	Local	Planted Area (ha)		Production		Unity
		2001	2002*	2001	2002*	
<i>Orange</i>	<i>Brazil</i>	825.228	828.437	16.983.436	18.931.919	tons
	<i>Center</i>	9.289	9.921	131.289	145.866	
	<i>North</i>	18.280	16.724	252.317	233.539	
	<i>North-East</i>	109.584	111.233	1.530.322	1.731.698	
	<i>South</i>	52.003	49.210	795.326	740.559	
	<i>South-East</i>	636.072	641.349	14.250.578	16.080.257	
	<i>Espirito Santo</i>	2.735	2.752	29.343	29.907	
	<i>Minas Gerais</i>	43.895	43.418	575.590	599.999	
	<i>Rio de Janeiro</i>	7.955	7.121	115.753	104.501	
	<i>São Paulo</i>	581.487	588.058	13.529.892	15.345.850	
<i>Banana</i>	<i>Brazil</i>	510.313	523.757	6.177.293	6.455.067	tons
<i>Coconut</i>	<i>Brazil</i>	275.551	273.306	1.420.547	1.811.773	10^3 fruits
<i>Pineapple</i>	<i>Brazil</i>	63.282	64.150	1.468.897	1.450.033	10^3 fruits
<i>Papaya</i>	<i>Brazil</i>	30.733	31.080	722.986	857.824	tons

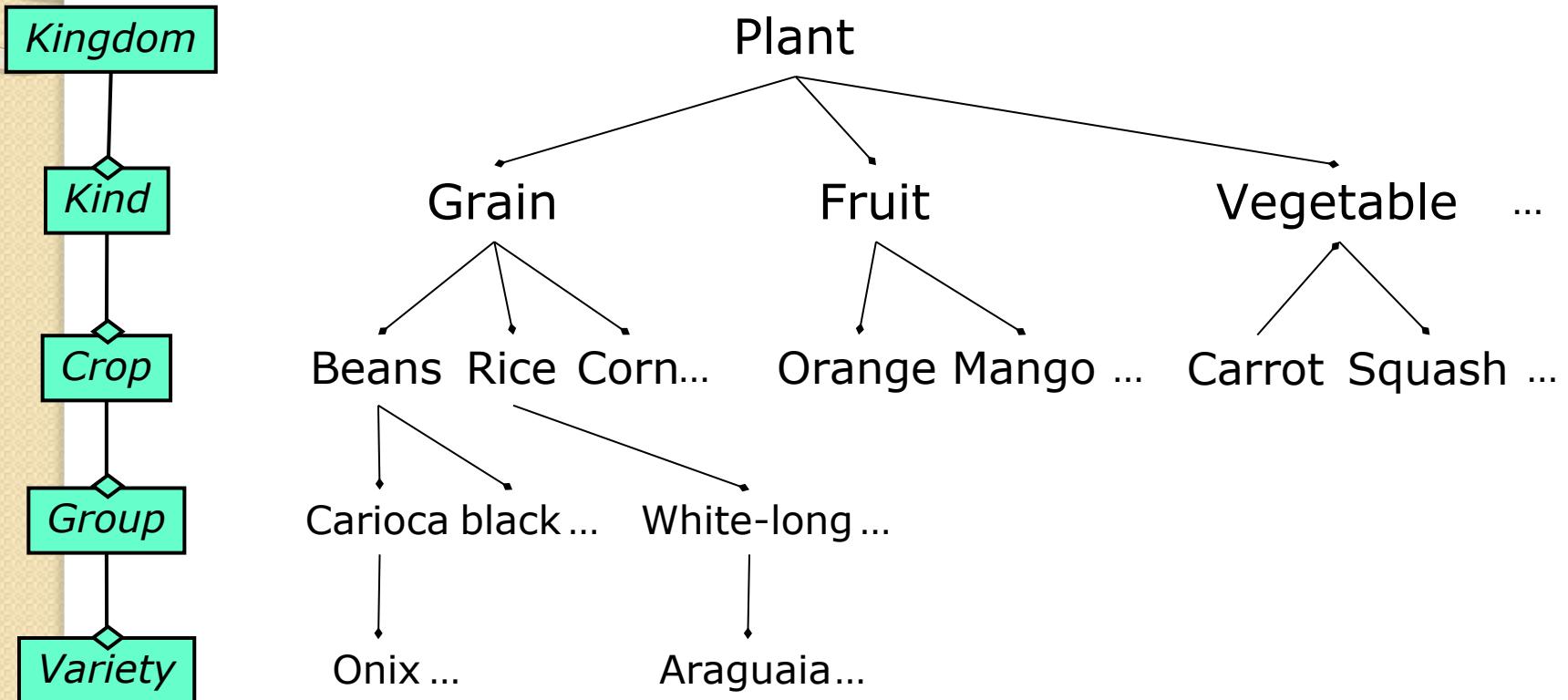
* Estimation by December 2002

DW Produção Agrícola

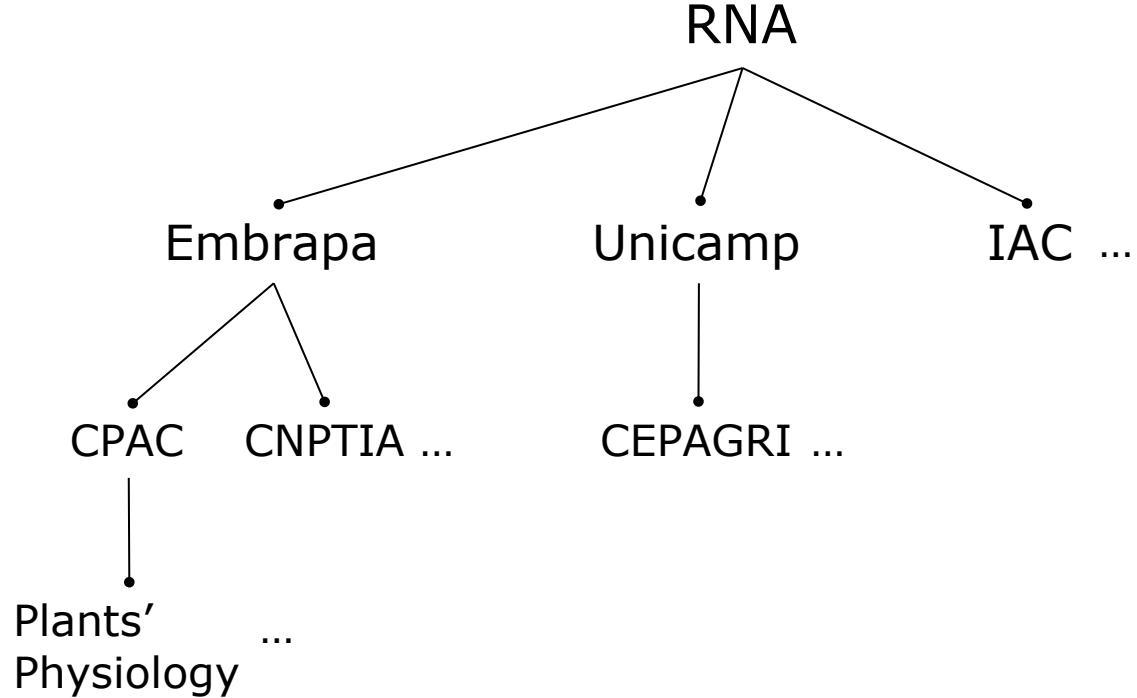
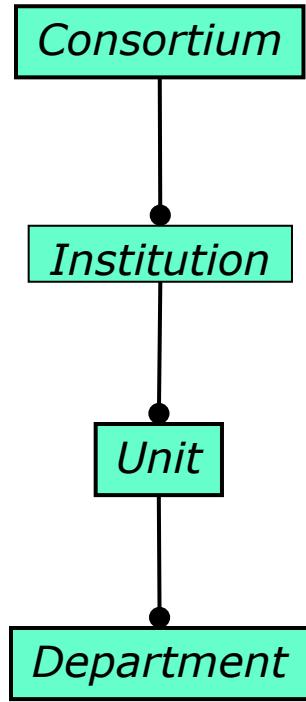
(Renato Deggau, Epagri/UFSC, 2009)

Produtos	Tempo	Local	Soma Área Plantada
+ All Produtos	+ All Tempo	+ All Local	63.682
Cebola	2007	+ Norte	0
		+ Nordeste	15.934
		+ Centro Oeste	1.348
		+ Sudeste	7.788
		+ MG	1.534
		+ SP	6.125
		+ RJ	0
		+ ES	129
		+ Sul	38.612
		+ PR	6.653
		+ SC	20.795
		+ Oeste Catarinense	2.135
		+ Norte Catarinense	412
		+ Serrana	2.082
		+ Vale do Itajaí	9.689
		+ Grande Florianópolis	6.381
		+ Sul Catarinense	96
		+ RS	11.164

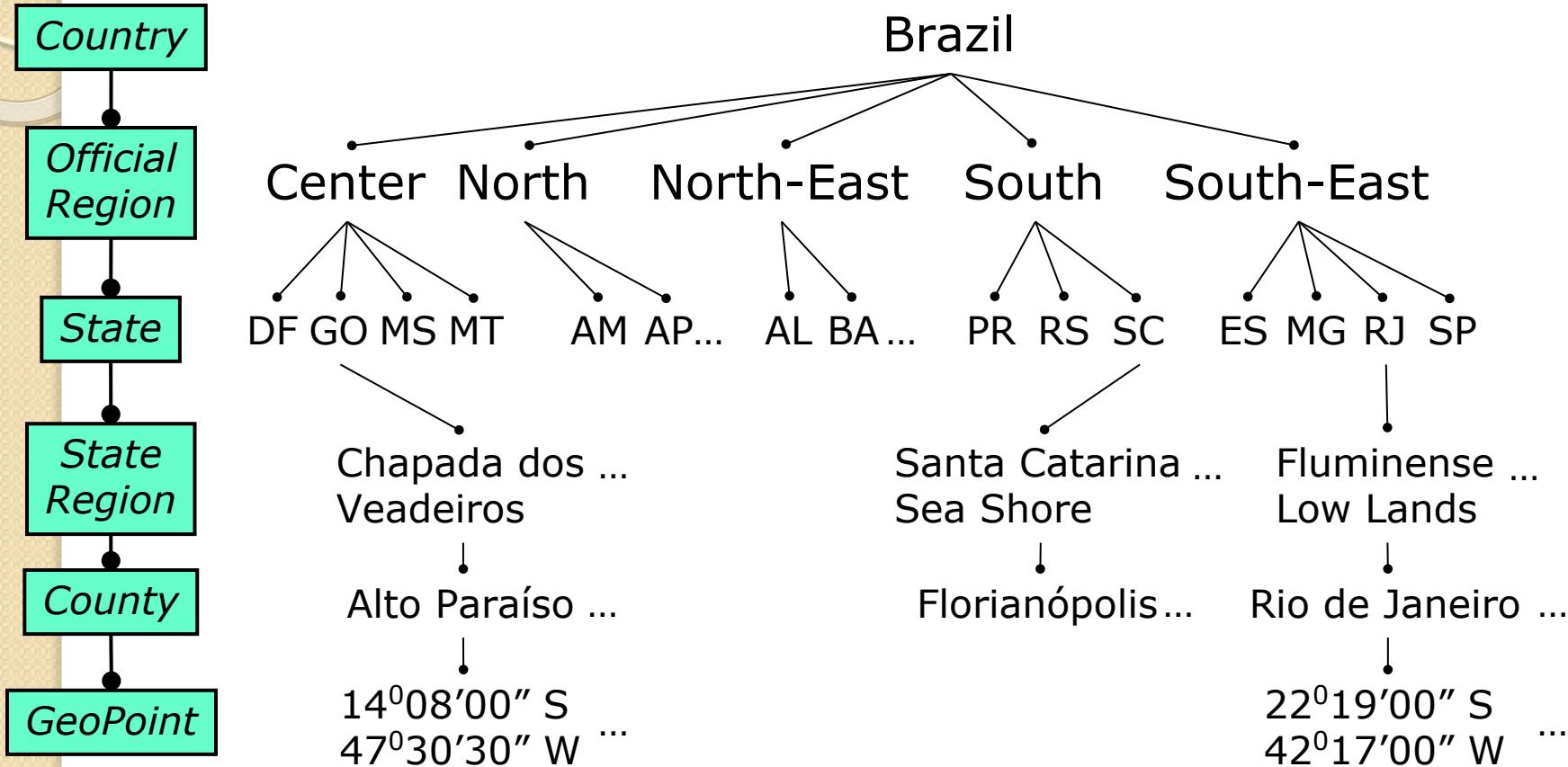
The Agricultural Product Dimension



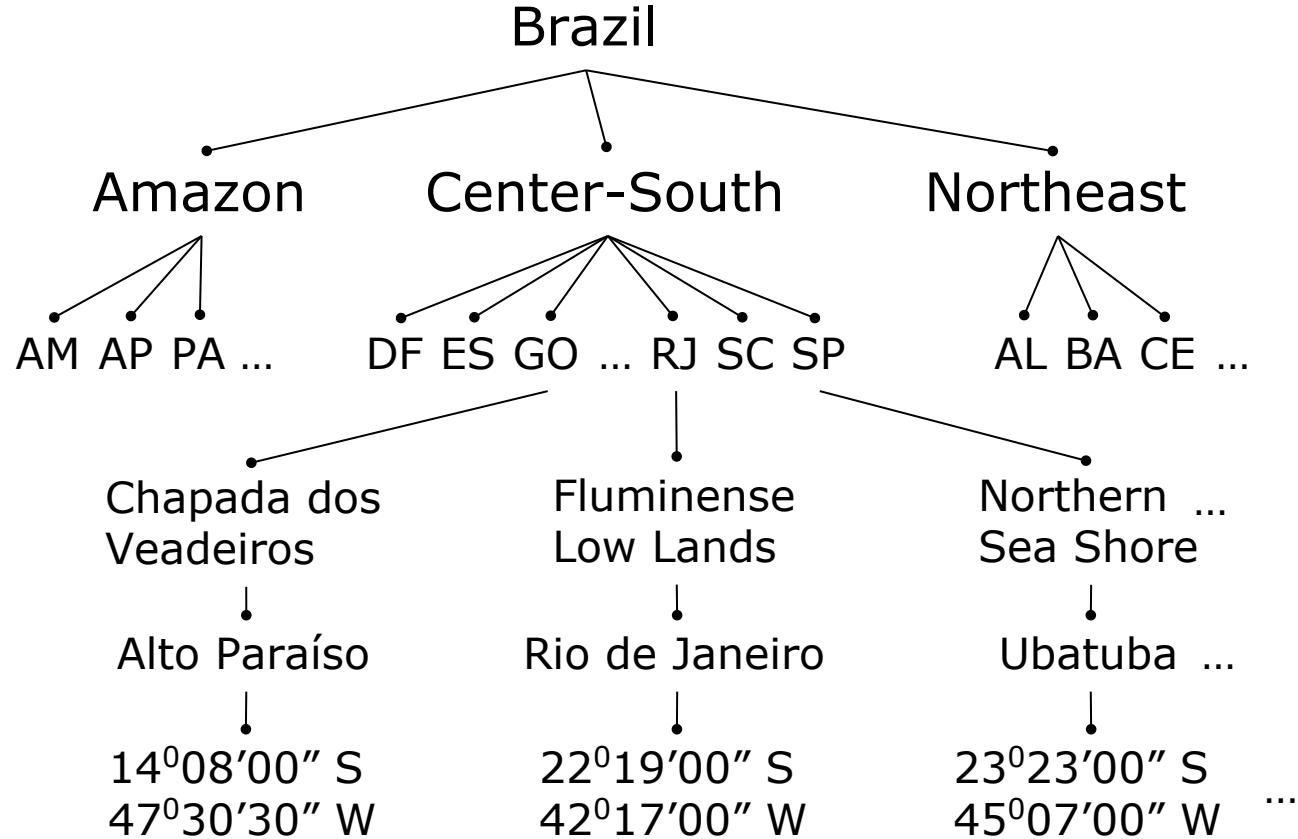
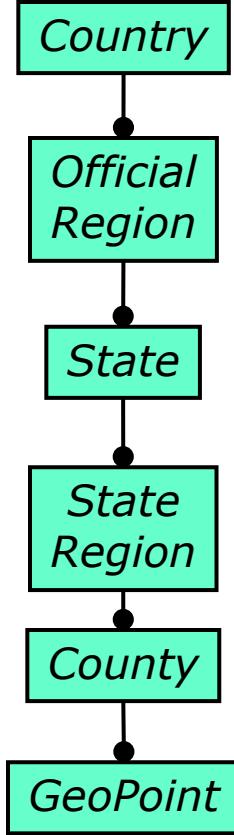
The Organizations Dimension



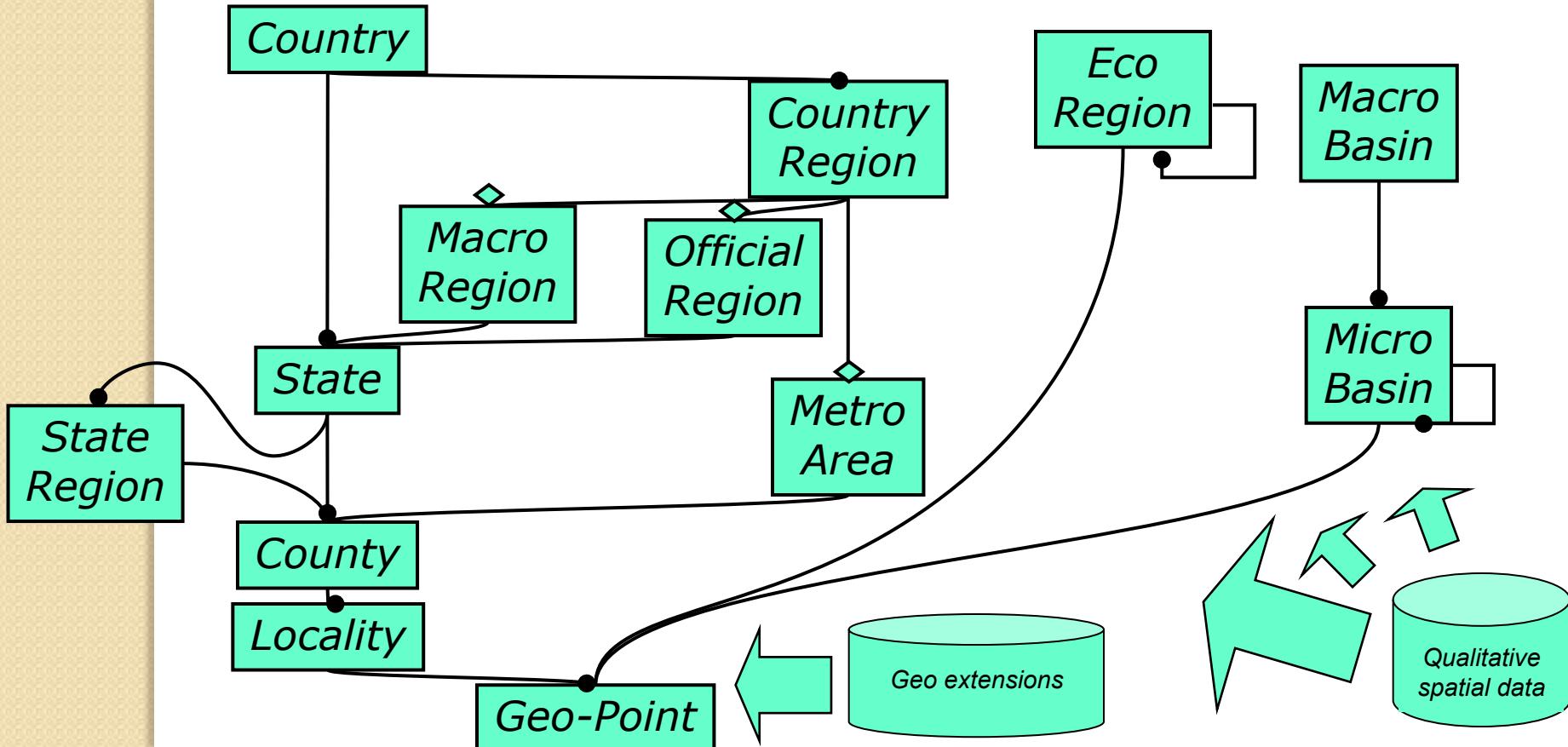
Instances (members) of Territory (I)



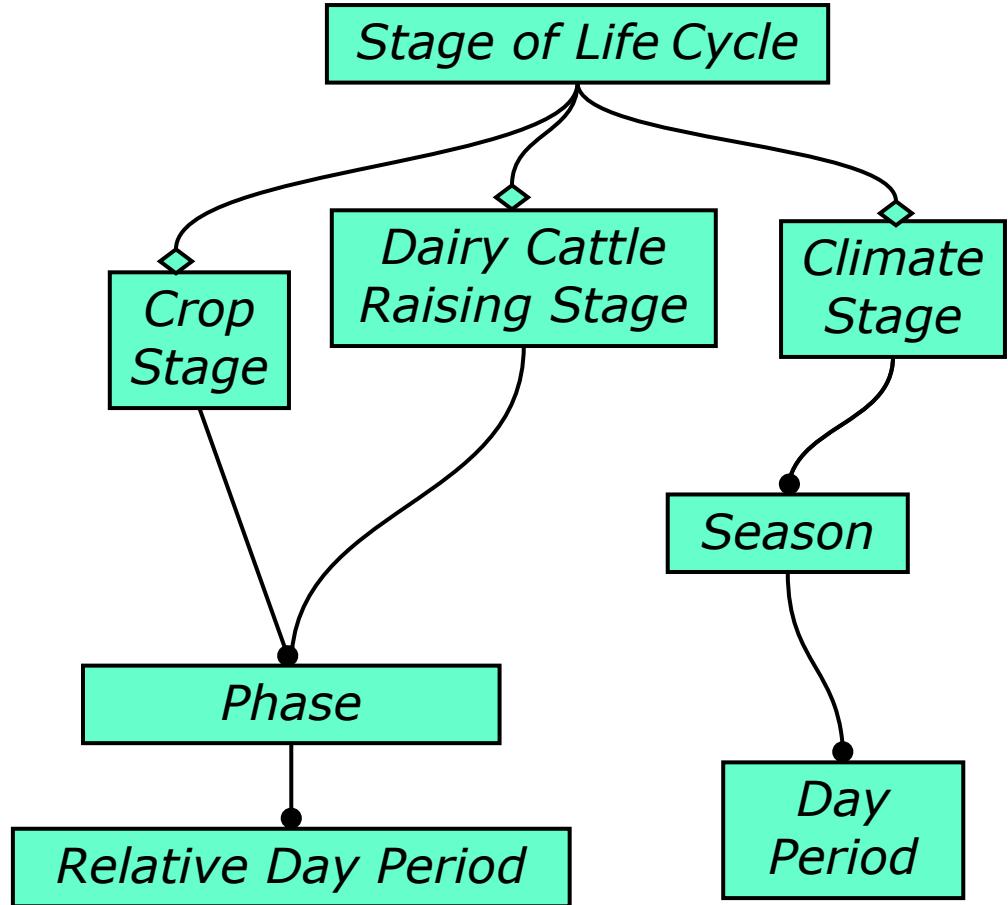
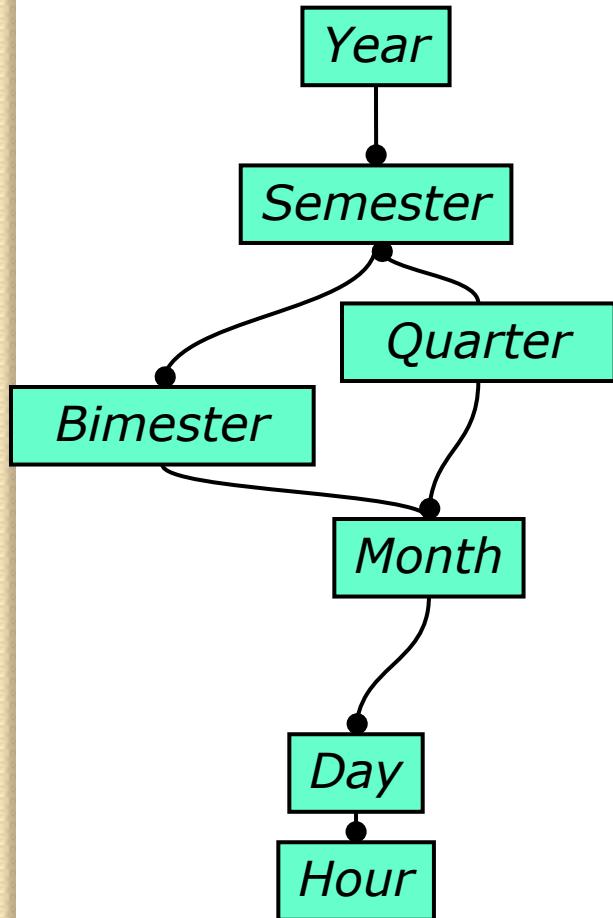
Instances (members) of Territory (II)



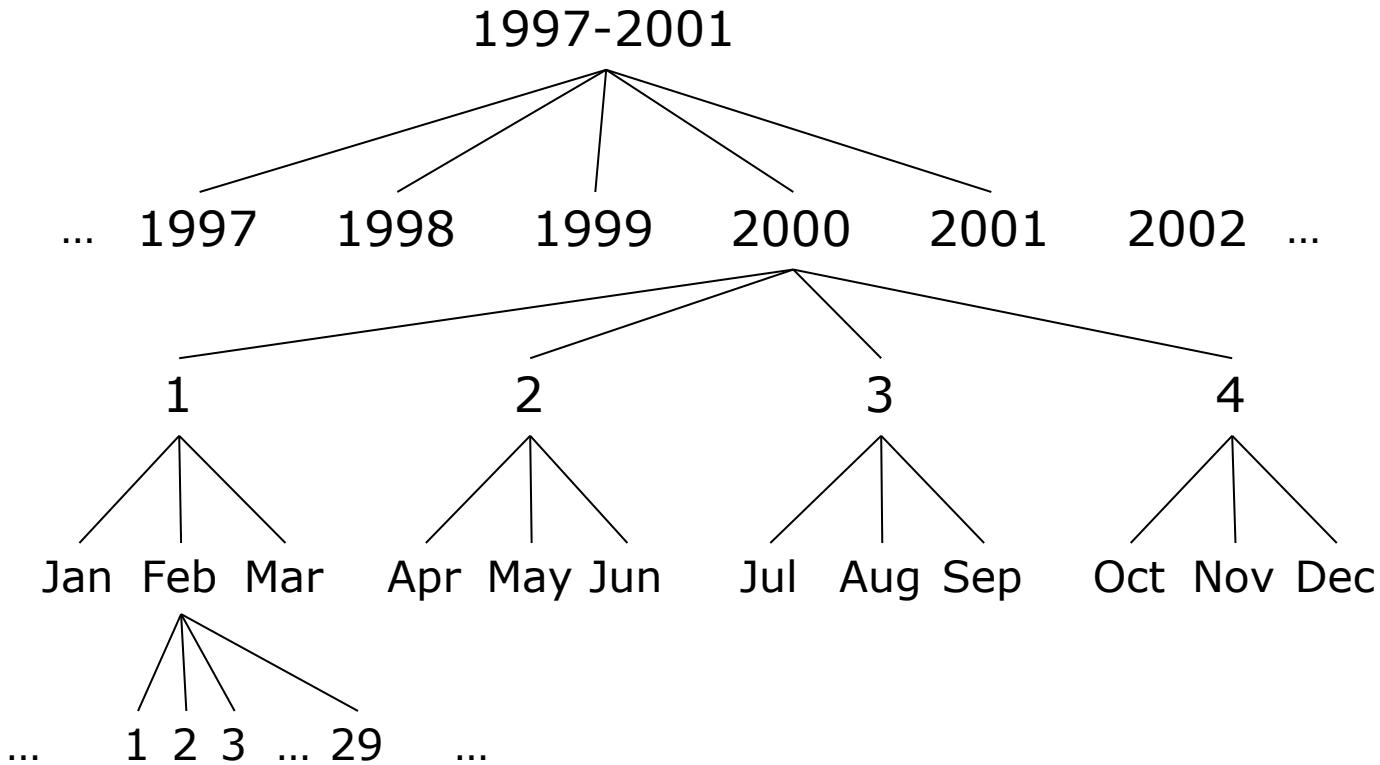
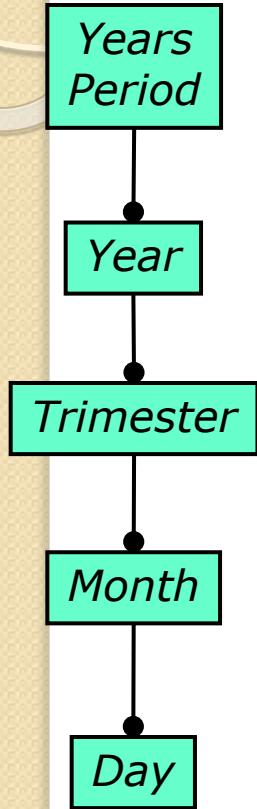
The Territory Dimension



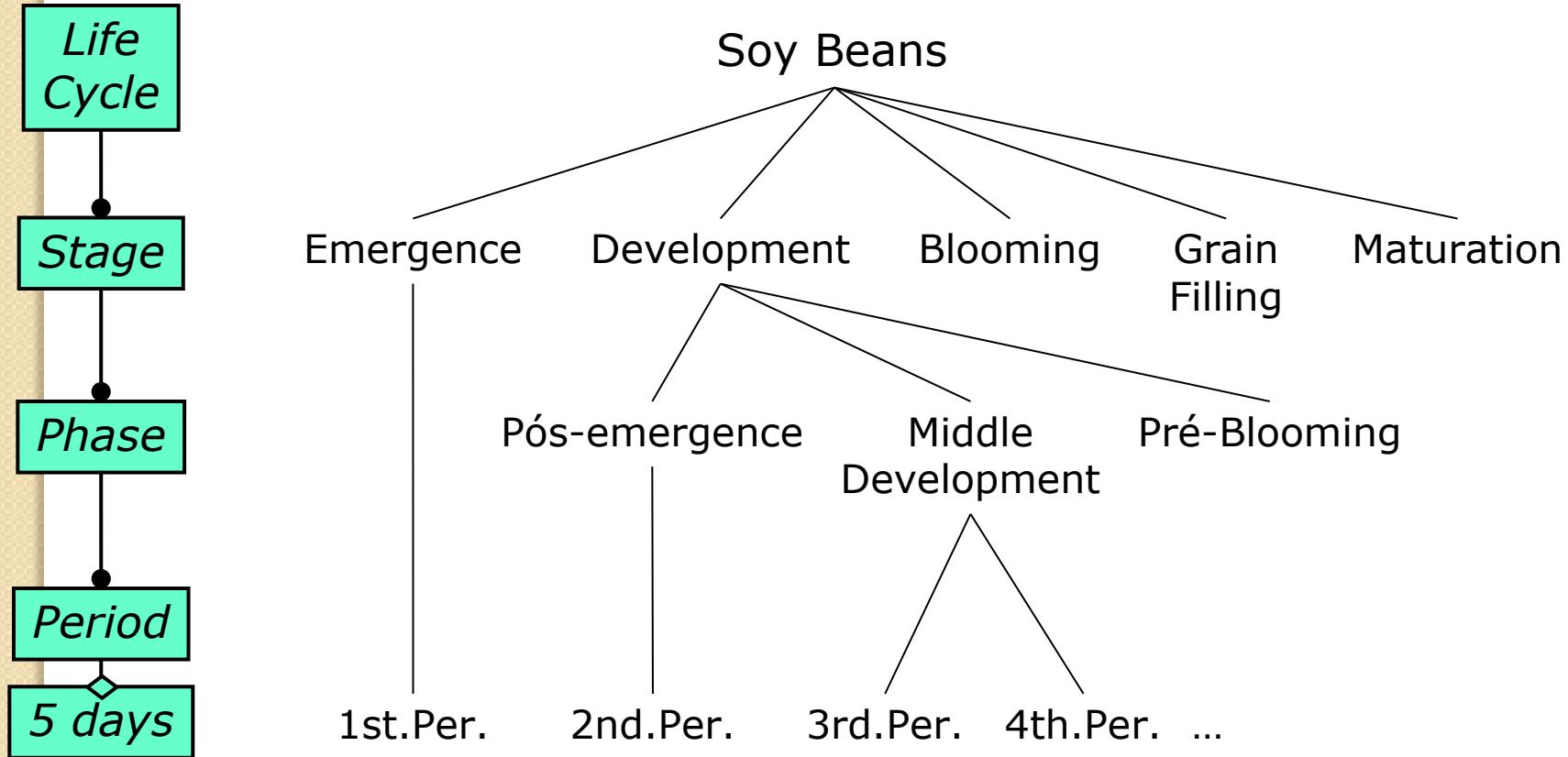
The Time Dimension



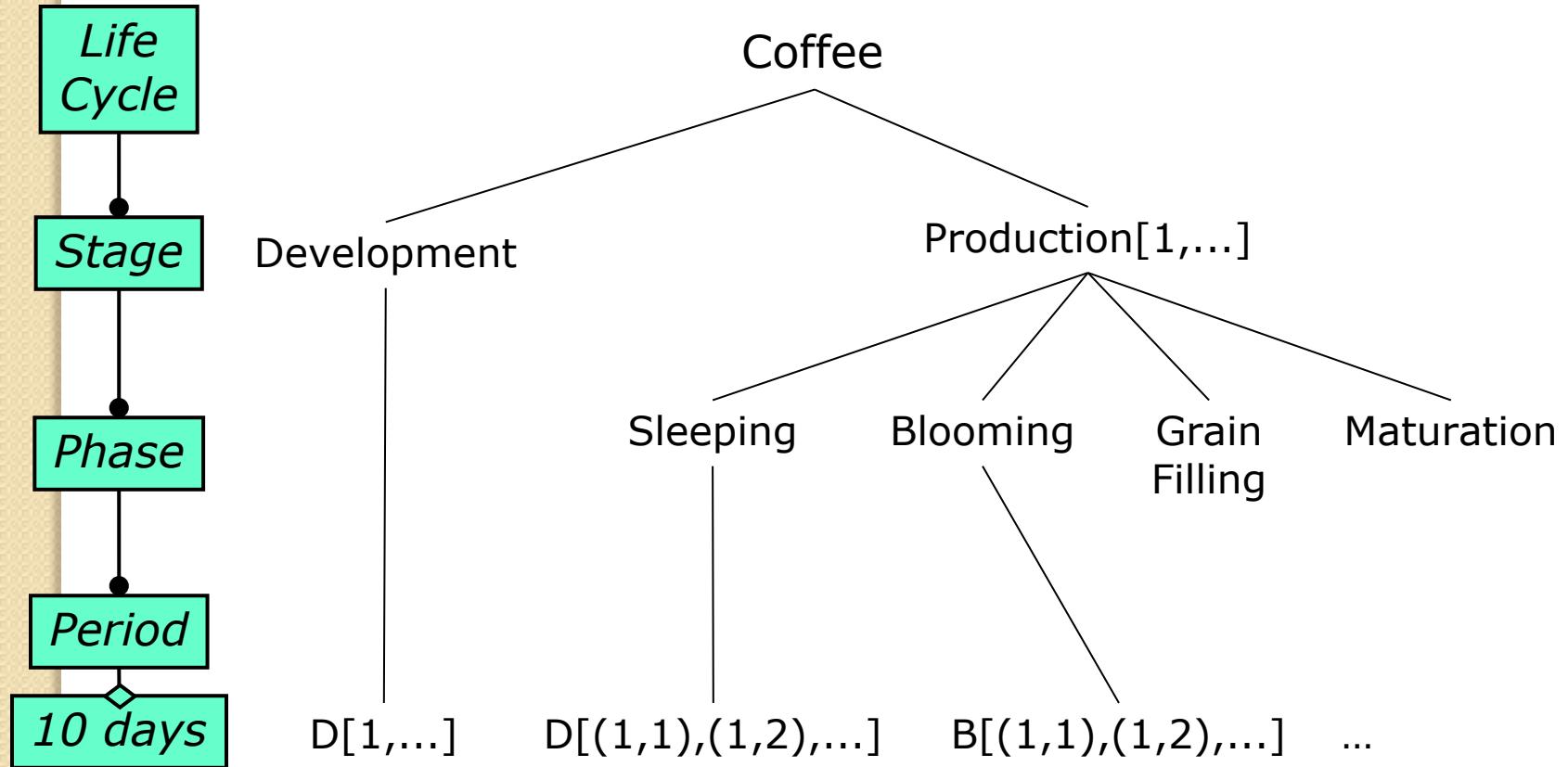
Calendar Time



Stages of an Annual Crop



Stages of a Permanent Crop



Normalização

Tabelas de Fatos:

Fato VENDAS

chave_tempo
chave_produto
chave_mercado
total_venda
unidades
custo

Tabelas de Dimensão:

DIMENSÃO PRODUTO

CHAVE_PRODUTO
DESCRIÇÃO
MARCA
CATEGORIA

Dimensão tempo não normalizada

Id Tempo	Dia	Mês	Ano	Estação	Dia Útil?	Dia Semana	Semestre
991020	20	10	1999	Primavera	S	Quarta	Segundo
991021	21	10	1999	Primavera	S	Quinta	Segundo
991022	22	10	1999	Verão	N	Sexta	Segundo
991023	23	10	1999	Verão	N	Sábado	Segundo

DW com dimensão loja



Disposição dos dados

Dimensão Produto

Chave Produto	Descrição	ID Produto	Departamento	Subcategoria
1	Guaraná	900087	Bebidas	Refrigerante
2	Soda	900088	Bebidas	Refrigerante

Dimensão Loja

Chave Loja	Nome Loja	Cidade
10	Beira Mar	Florianópolis
20	Coqueiros	Florianópolis

Dimensão Tempo

Chave Tempo	Dia	Mês	Ano	Feriado
100	01	05	2019	Dia do Trabalho
101	02	05	2019	

Fato Vendas

Chave Tempo	Chave Produto	Chave Loja	Total Vendas	Total Unidades
100	1	10	2400,00	2.000
100	1	20	1800,00	1.500
100	2	10	1000,00	1.000
100	2	20	1500,00	1.500

Dimensionamento do BD

- **Dimensão Tempo:** 2 anos x 365 dias = 730 dias.
- **Dimensão Produto:** 30.000 produtos sendo 3.000 vendidos todos os dias.
- **Dimensão Loja:** 300 lojas
- **Número de registros de fatos básicos** (menor grão) = $730 \times 3.000 \times 300 = 657$ milhões de registros
- **Número de campos** = 3 chaves + 4 fatos = 7
- **Tamanho básico da tabela de fatos** = $657\text{ milhões} \times 7\text{ campos} \times 4\text{ bytes} = 18\text{ GB}$

Mudanças em dimensões (Slowly Changing Dimensions - SCD)

Eventualmente, dimensões (especialmente as grandes como Produto e Cliente) podem sofrer alterações...

Pergunta: Como tratar estas modificações ao longo do tempo?

Resposta: Há várias técnicas para tratar Dimensões de Modificação Lenta (SCD)

DIMENSÃO	PRODUTO
COD_PRODUTO	
DESCRIÇÃO	
NUMERO_ID	
DEPARTAMENTO	
CATEGORIA	
SUBCATEGORIA	
MARCA	
TAMANHO_EMB	
TIPO_EMB	
PESO	
...	

SCD - Técnica I- sobrescrita

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Cia XYZ	SC

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Cia XYZ	SP

SCD - Técnica 2- inserir nova linha (membro) na dimensão

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Begin	new
123	ABC	Cia XYZ	SC	2004-12-22T00:00:00	2017-12-22T23:59:59
123	ABC	Cia XYZ	SP	2018-01-01T00:00:00	NULL

SCD - Técnica 3 - inserir nova coluna (atributo) na dimensão

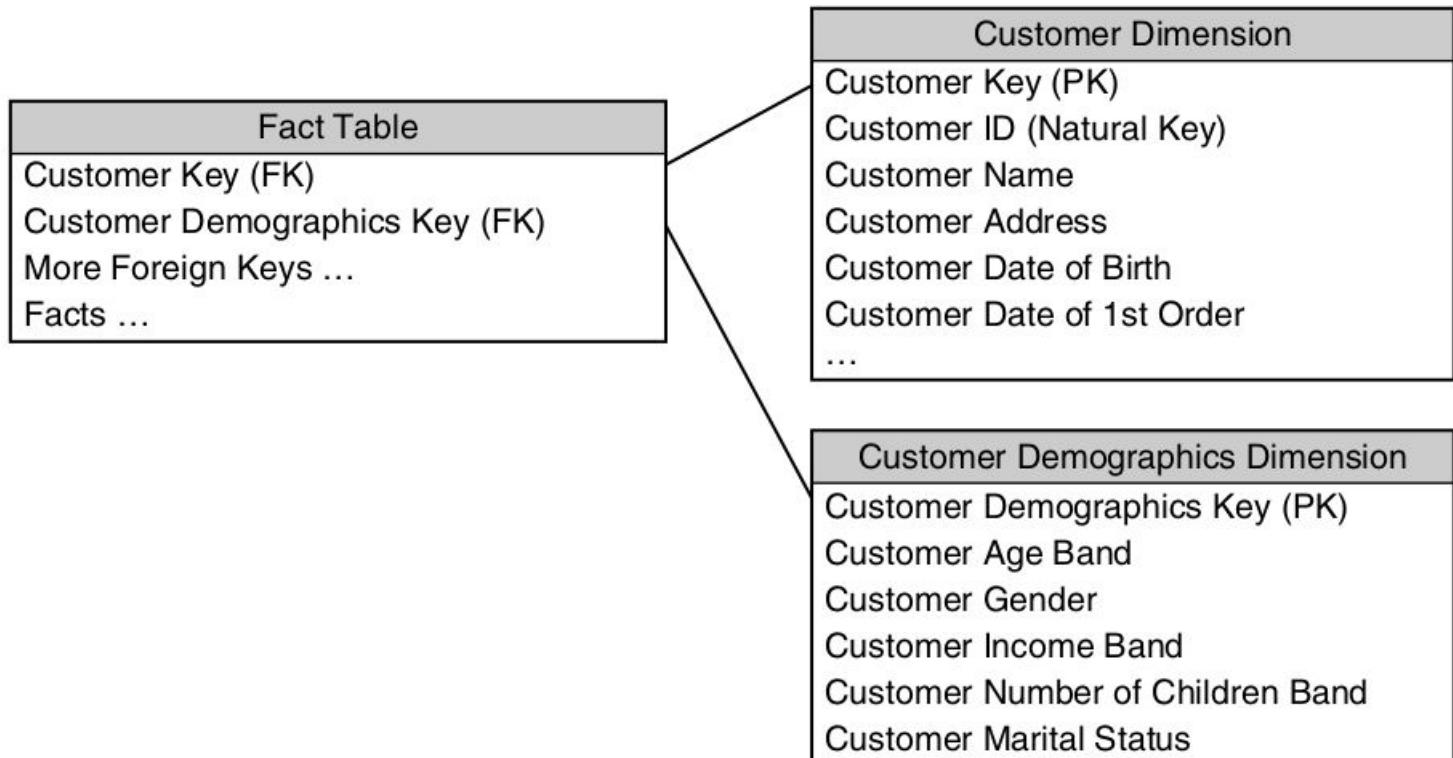
Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Effect	New_Supplier_State
123	ABC	Cia XYZ	SC	2004-12-22T00:00:00	SP

SCD - Técnica 3

representação híbrida

PRODUCT KEY	Product Description	Product Code	Historical Department	Current Department	Effective Date	Expiration Date	Current Row Ind
12345	IntelliKidz 1.0	ABC999-Z	Education	Critical Thinking	2/15/2007	5/31/2007	Not current
25984	IntelliKidz 1.0	ABC999-Z	Strategy	Critical Thinking	6/1/2007	12/31/2007	Not current
34317	IntelliKidz 1.0	ABC999-Z	Critical Thinking	Critical Thinking	1/1/2008	12/31/9999	Current

Mini-dimensões



Muitas dimensões

(só indicadas - esboço no início da modelagem)

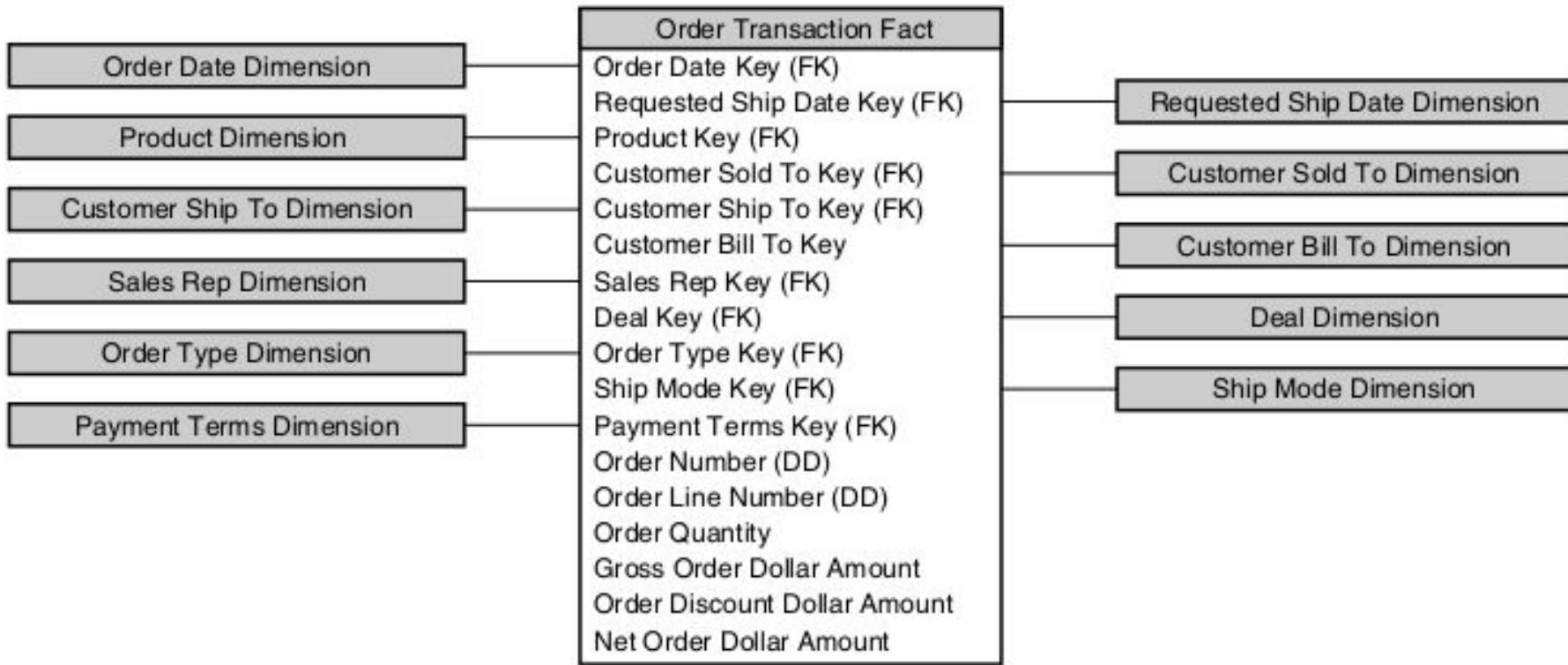
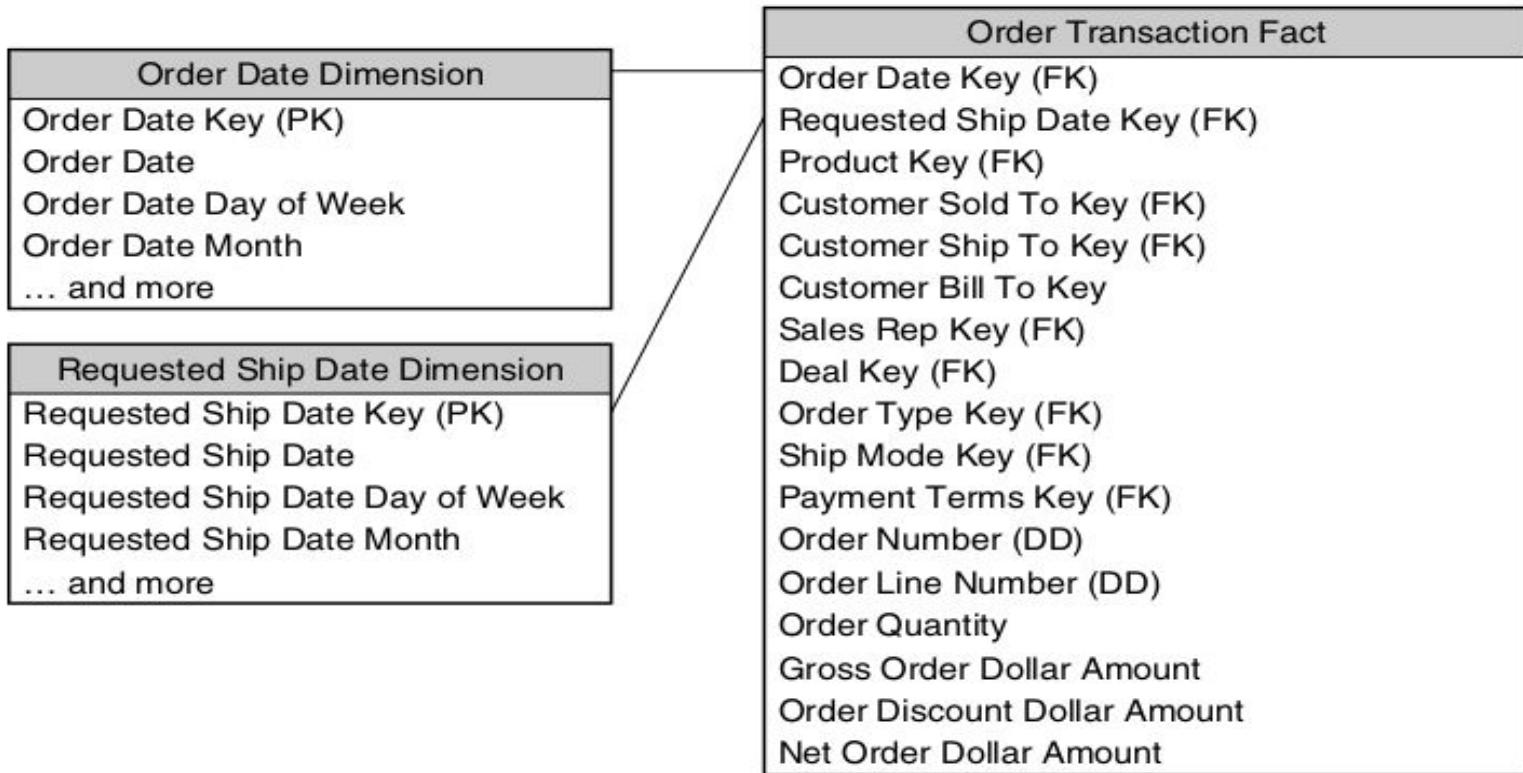


Tabela fato com múltiplos fusos horários



Dimensões data com papéis



Tabelas fato com granularidades distintas em dimensões análogas

Sales Fact Table
Date Key (FK)
Product Key (FK)
More Foreign Keys ...
Sales Quantity
Sales \$ Amount

Product Dimension
Product Key (PK)
Product Description
SKU Number (Natural Key)
Brand Description
Subclass Description
Class Description
Department Description
Color
Size
Display Type

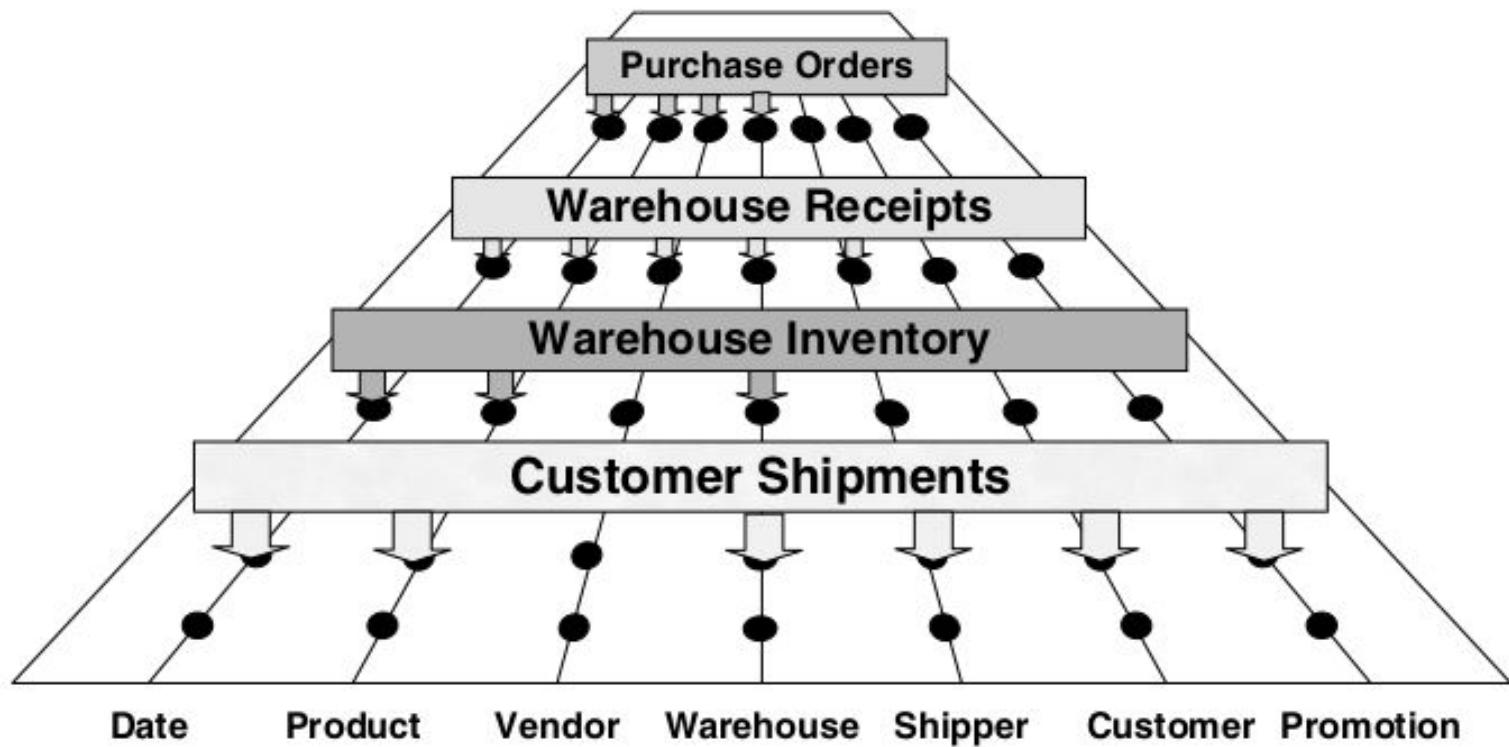
Sales Forecast Fact Table
Month Key (FK)
Brand Key (FK)
More Foreign Keys ...
Forecast Quantity
Forecast \$ Amount

Brand Dimension
Brand Key (PK)
Brand Description
Subclass Description
Class Description
Department Description
Display Type

Junk dimension - termos muito técnicos e específicos de domínio

Invoice Indicator Key	Payment Terms	Order Mode	Ship Mode
1	Net 10	Telephone	Freight
2	Net 10	Telephone	Air
3	Net 10	Fax	Freight
4	Net 10	Fax	Air
5	Net 10	Web	Freight
6	Net 10	Web	Air
7	Net 15	Telephone	Freight
8	Net 15	Telephone	Air
9	Net 15	Fax	Freight
10	Net 15	Fax	Air
11	Net 15	Web	Freight
12	Net 15	Web	Air
13	Net 30	Telephone	Freight
14	Net 30	Telephone	Air
15	Net 30	Fax	Freight
16	Net 30	Fax	Air
17	Net 30	Web	Freight
18	Net 30	Web	Air
19	Net 45	Telephone	Freight
20	Net 45	Telephone	Air
21	Net 45	Fax	Freight
22	Net 45	Fax	Air
23	Net 45	Web	Freight
24	Net 45	Web	Air

DW bus (dimensões compartilhadas por DMs)



Bus matrix

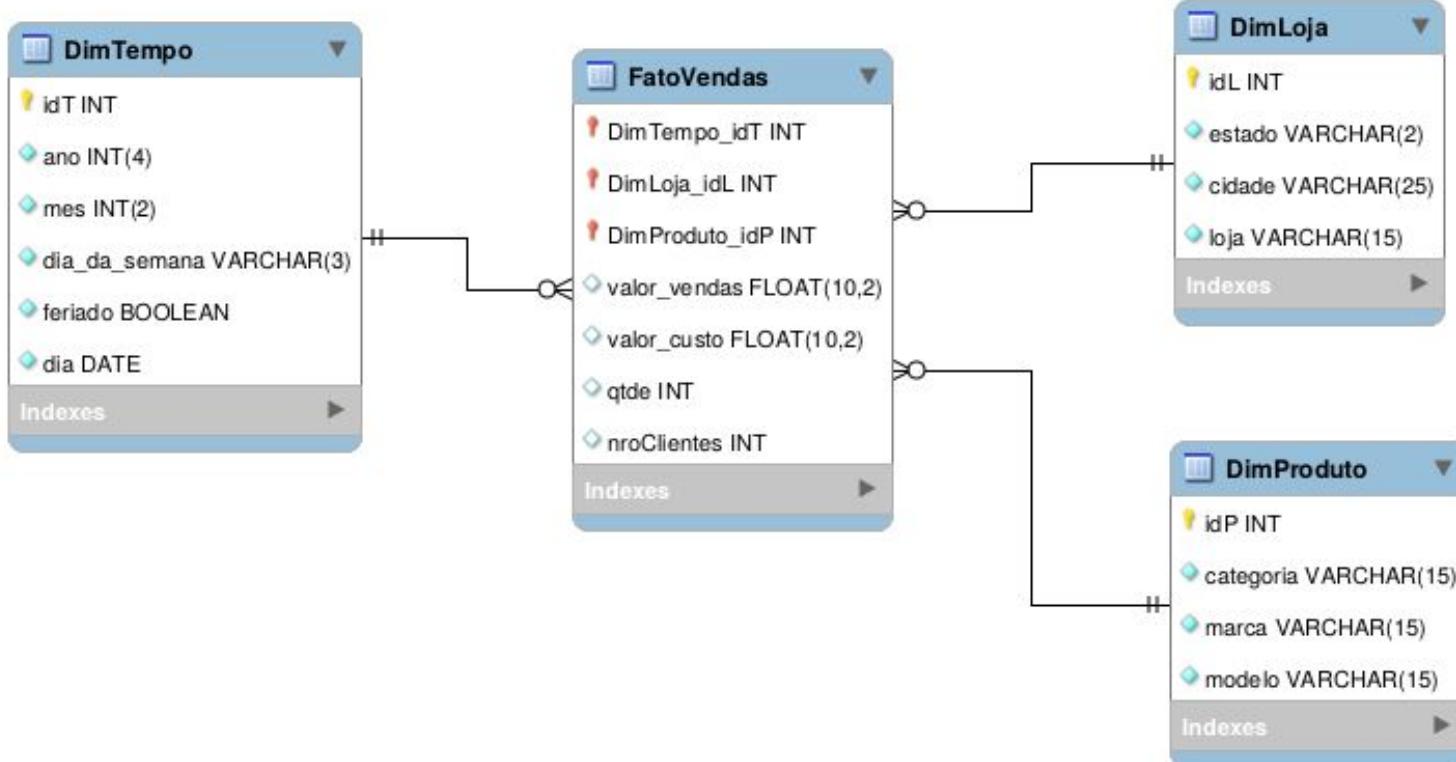
	<i>Date</i>	<i>Raw Material</i>	<i>Supplier</i>	<i>Plant</i>	<i>Product</i>	<i>Shipper</i>	<i>Warehouse</i>	<i>Customer</i>	<i>Sales Rep</i>	<i>Promotion Deal</i>
Raw Material Purchasing	X	X	X	X		X				
Raw Material Delivery	X	X	X	X		X				
Raw Material Inventory	X	X	X	X						
Bill of Materials	X	X		X	X					
Manufacturing	X	X	X	X	X					
Shipping to Warehouse	X			X	X	X	X			
Finished Goods Inventory	X			X			X			
Customer Orders	X				X	X		X	X	X
Shipping to Customer	X				X	X	X	X	X	X
Invoicing	X				X		X	X	X	X
Payments	X				X			X	X	X
Returns	X				X	X		X	X	X

Entregáveis do projeto lógico

- Esquema lógico de alto nível
- Dicionário de dados: lista atributos e métricas, além de descrever seus tipos e significados
- Planilha do projeto dimensional detalhado para cada tabela do esquema
- Lista de dificuldades, desafios e riscos

Exemplo de esquema lógico

(produzido usando o MySQL Workbench)



Exemplo de planilha de projeto dimensional detalhado

Table Name:	DimOrderInfo
Table Type	Dimension
View Name	OrderInfo
Description	OrderInfo is the "junk" dimension that includes miscellaneous information about the Order transaction
Used in schemas	Orders
Generate script?	Y

Column Name	Target										Source						Comments
	Description	Datatype	Size	Key?	FK To	NULL?	Default Value	Unknown Member	Example Values	SCD Type	Source System	Source Schema	Source Table	Source Field Name	Source Datatype	ETL Rules	
OrderInfoKey	Surrogate primary key	smallint		PK	ID	N		-1	1, 2, 3, 4...		ETL Process						Standard surrogate key
BKSalesReasonID	Sales reason ID from source system	smallint				N		-1			OEI	Sales	SalesReason	SalesReasonID	int		Convert to char; left-pad with zero. R for reseller row.
Channel	Sales channel	char	8					Unknown	Reseller, Internet, Field Sales	1	OEI	Sales	SalesReason	Derived			"Internet" for real sales reasons, "Reseller" for reseller row.
SalesReason	Reason for the sale, as reported by the customer	varchar	30					Unknown		1	OEI	Sales	SalesReason	Name	nvarchar(50)		Convert to varchar; "Reseller" for reseller row.
SalesReasonType	Type of sales reason	char	10					Unknown	Marketing, Promotion, Other	1	OEI	Sales	SalesReason	ReasonType	nvarchar(50)		Convert to varchar; "Reseller" for reseller row.
AuditKey	What process loaded this row?	int		FK	Audit Dim	N		-1		1	Derived						Populated by ETL system using standard technique

Comments

Order_Info is a "junk" dimension with only a handful of rows based on "Channel" and "Sales Reason". We currently have only three channels and sales reasons only for field sales and Internet sales. We can eliminate a dimension by combining these two.

Bus matrix detalhada

Business Process	Fact Tables	Granularity	Facts	Date	Policyholder	Coverage	Covered Item	Employee	Policy	Claim	Claimant	3rd Party
Policy Transactions	Corporate Policy Transactions	1 row for every policy transaction	Policy Transaction Amount	X Trxn Eff	X	X	X	X	X			
	Auto Policy Transactions	1 row per auto policy transaction	Policy Transaction Amount	X Trxn Eff	X	X Auto	X Auto	X	X			
	Home Policy Transactions	1 row per home policy transaction	Policy Transaction Amount	X Trxn Eff	X	X Home	X Home	X	X			
Policy Premium Snapshot	Corporate Policy Premiums	1 row for every policy, covered item, and coverage each month	Written Premium Revenue Amount, Earned Premium Revenue Amount	X	X	X	X	X Agent	X			
	Auto Policy Premiums	1 row per auto policy, covered item, and coverage each month	Written Premium Revenue Amount, Earned Premium Revenue Amount	X	X	X Auto	X Auto	X Agent	X			
	Home Policy Premiums	1 row per home policy, covered item, and coverage each month	Written Premium Revenue Amount, Earned Premium Revenue Amount	X	X	X Home	X Home	X Agent	X			
Claim Transactions	Claim Transactions	1 row for every claim transaction	Claim Transaction Amount	X Trxn Eff	X	X	X	X	X	X	X	X
	Claim Accumulating Snapshot	1 row per covered item and coverage on a claim	Original Reserve Amount, Assessed Damage Amount, Reserve Adjustment Amount, Current Reserve Amount, Open Reserve Amount, Claim Amount Paid, Payments Received, Salvage Received, Number of Transactions	X	X	X	X	X Agent	X	X	X	
	Accident Event	1 row per loss party and affiliation in an auto claim	Implied Accident Count	X	X	X Auto	X Auto		X	X Auto	X	

Descrição de fontes de dados

Source	Business Owner	IS Owner	Platform	Location	Description
Gemini	Tom Owens	Alison Jones	Unix	HQ - Chicago	Distribution center inventory
Billings	Craig Bennet	Steve Dill	MVS	MF - Dallas	Customer Billings
Plant	Sylvia York	Bob Mitchell	Unix	6 Plants across country	Plant shipments
Sales Forecast	Sandra Phillips	None	Windows	HQ - Sales Dept	Spreadsheet-based consolidated sales forecast
Competitor Sales	Sandra Phillips	None	Windows	HQ - Sales Dept	Access database containing data from external supplier

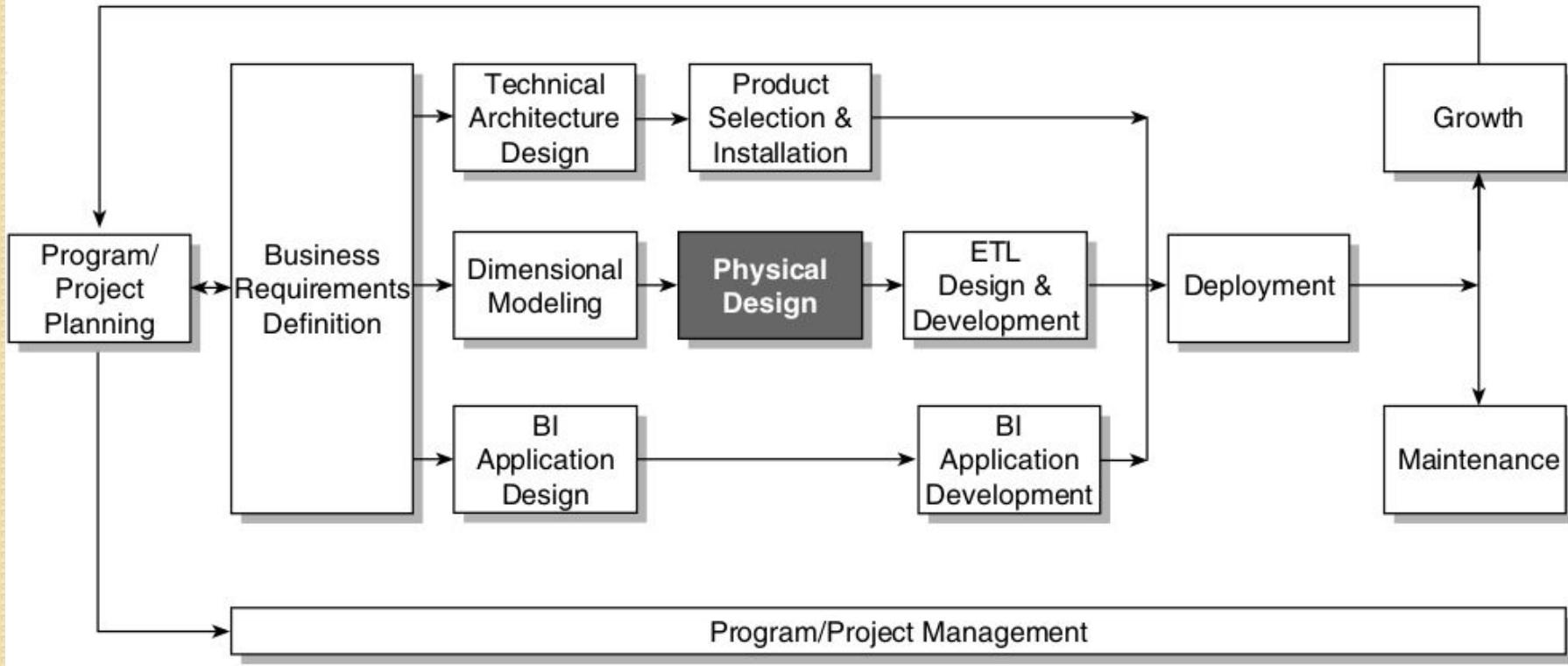
Derivação de fatos

<i>Fact Group</i>	<i>Measure Name</i>	<i>Measure Description</i>	<i>Agg Rule</i>	<i>Formula</i>	<i>Constraints</i>	<i>Transformations</i>
POS	\$ Sales	The dollar amount of the goods sold through the retail channel.	Sum	Sum(Dollar Sales)	None	None
POS	Total US \$ Sales	The dollar amount sold for the total US geography.	Sum	Sum(Dollar Sales)	Geography= Total US	None
POS	% of Total US \$	Dollar sales as a percentage of total US geography.	Recalc	$(\$ \text{Sales}/\text{Total US } \$ \text{Sales}) * 100$	NA	NA
POS	Prev. \$ Sales	The dollar amount of the goods sold through the retail channel during the previous period.	Sum	Sum(Dollar Sales)	None	Previous period
POS	Prev. Tot US \$ Sales	The dollar amount sold for the total US during the previous period.	Sum	Sum(Dollar Sales)	Geography= Total US	Previous period
POS	Prev. % of Total US \$	The previous period dollars as a percentage of the previous period total US dollars.	Recalc	$(\text{Prev } \$ \text{Sales}/\text{Prev Tot US } \$ \text{Sales}) * 100$	NA	NA
POS	\$ Chg vs. Prev	The actual change in dollars from previous period.	Sum	$\$ \text{Sales} - \text{Prev } \$ \text{Sales}$	NA	NA
POS	Units	The number of consumer units sold.	Sum	Sum(Units)	None	None
POS	Avg Retail Price	The average price at the register.	Recalc	$\$ \text{Sales}/\text{Units}$	None	None
Inv	Inventory \$	The dollar value of units in inventory.	Sum w/Limit	Sum(inv Units) expect across time, then take value from max date.		
Fcst	Forecast \$	The dollar amount of the expected sales through the retail sales channel.	Sum	Sum(Forecast Dollars)	None	None
Multi Group	% Var to Forecast \$	The percentage difference between actual and forecast sales dollars.	Recalc	$(\text{Sum}(\text{Dollar Sales}) - \text{Sum}(\text{Forecast Dollars})) / \text{Sum}(\text{Dollar Sales})$	None	None

Tópicos

- I. O Modelo de dados dimensional**
 - Fatos e dimensões
 - Esquemas em estrela, floco de neve ou hipercubos
 - Medidas de fatos e funções de agregação
 - Hierarquias, níveis e membros de dimensões
 - Operadores OLAP sobre esquemas dimensionais
- 2. Projeto de esquemas dimensionais em DWs**
 - Medidas de fatos e funções de agregação
 - Hierarquias, níveis e membros de dimensões
- 3. Projeto físico e de desempenho**
 - Padrões e esquema físico
 - Ajustes para eficiência: indexação, agregações, ...

Projeto físico no processo de DW



Projeto físico de DW

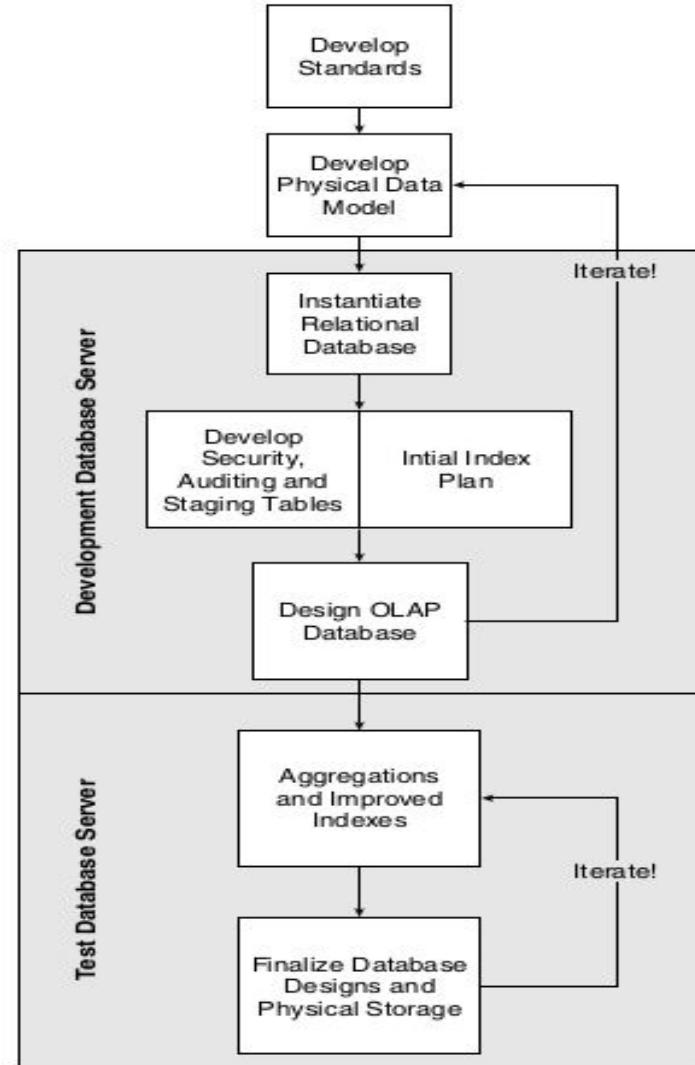
- Detalhes de implementação variam muito de acordo com o planejamento do projeto/programa de DW e com a arquitetura técnica.
- Ferramentas de SW e hardware evoluem rapidamente levando a diferentes possibilidades.

Projeto físico de DW - conselhos

- Faça um plano de implementação física e amarre-o ao plano geral do projeto/programa de DW.
- Não sucumba à tentação de fazer algo fácil (mas errado) como um paliativo temporário.
- Defina padrões e siga-os.
- Use as melhores ferramentas disponíveis e as mais apropriadas para o seu cenário.

Processo de alto nível para o projeto físico

(Kimball)

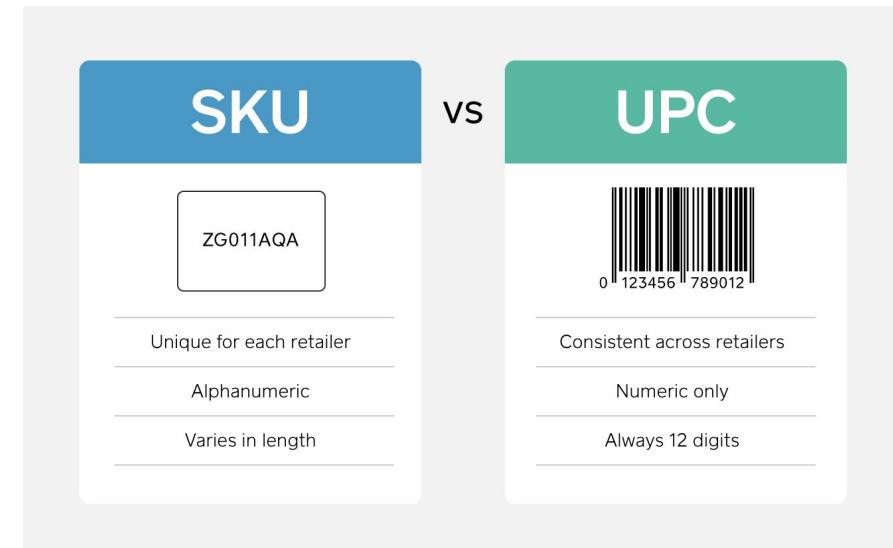


Uma dimensão não normalizada

Product	
PK	Product Key
	Product SKU
	Product Name
	Product Descr
	Product Color
	Product Subcategory Key
	Product Subcategory
	Product Subcategory Descr
	Product Category Key
	Product Category
	Product Category Descr
	other attributes...

SKU - Stock Keeping Unit

UPC - Universal Product Code



Uma dimensão em floco de neve

(menos eficiente porque requer junções)

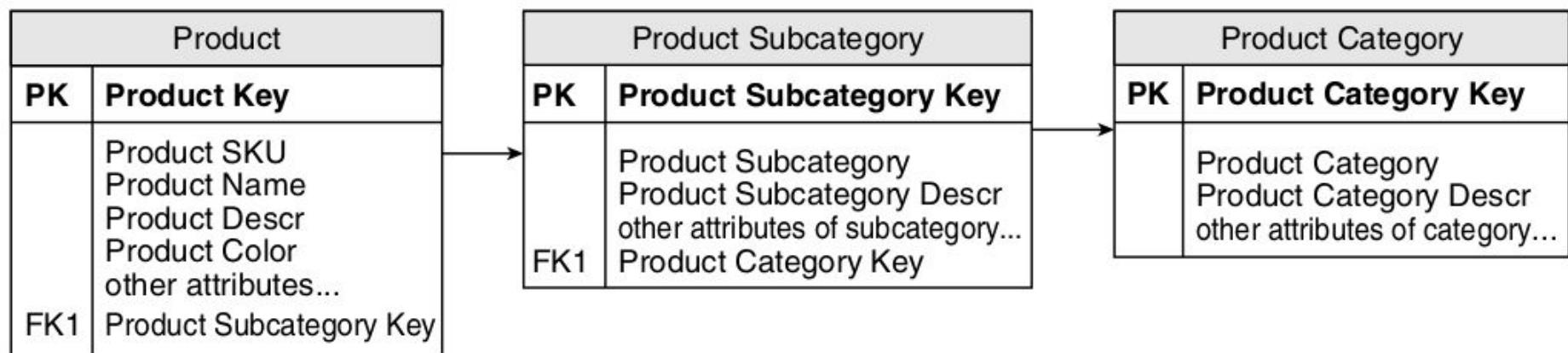
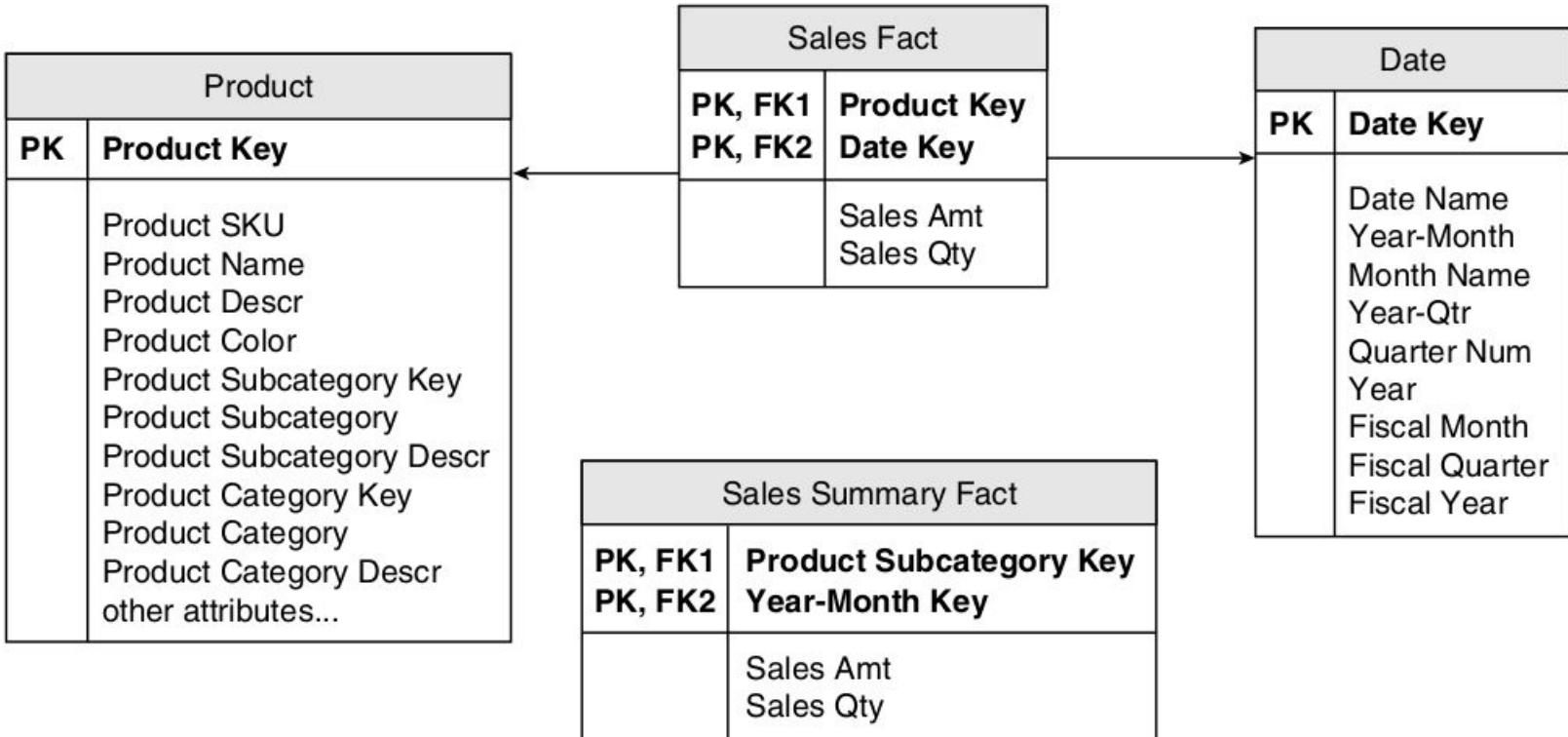


Tabela sumário



Entregáveis do projeto físico

- Esquema físico no servidor de desenvolvimento
- Mapeamento final fontes-destinos
- Plano inicial de indexação
- Projeto da base de dados OLAP
- Plano de (pré-)agregação
- Plano de particionamento (se necessário)

Trabalho de Modelagem

1. Planejamento
2. Definição dos requisitos e fontes de dados
3. Integração de dados
4. **Modelagem dimensional (será avaliada!)**
5. Projeto físico do banco de dados
6. Projeto das transformações de dados (ETC)