

Notas de Leitura em *Business Intelligence*

03 >> Modelação Dimensional de Dados

Orlando Belo

Departamento de Informática, Escola de Engenharia, Universidade do Minho
PORTUGAL



Resumo

A modelação dimensional de dados é uma das **atividades mais relevantes** de qualquer **projeto de *data warehousing***. Juntamente com o processo de levantamento e análise de requisitos, esta atividade contribui de forma muito significativa para o desenvolvimento do repositório de dados – *data warehouse* – de um sistema de *data warehousing*, uma vez que suporta todo **o processo de construção dos seus esquemas dimensionais de dados** subjacentes ao *data warehouse* em questão. A **modelação dimensional** permite a concepção das estruturas de dados de um *data warehouse* de acordo com as várias **perspectivas de análise dos agentes de decisão** do **domínio em questão** e de todos os processos de exploração de dados que estes lançam para satisfazer, por um lado, as suas necessidades mais básicas de *reporting* e, por outro, o cruzamento de dados entre uma ou mais estruturas de dados integradas no *data warehouse*.



Agenda

- Introdução.
- Desenvolvimento de sistemas de *data warehousing*.
- Construção de *data warehouses*.
- A adoção de um método.
- A área de suporte à decisão, o grão, os factos e sua representação.
- Dimensões e perspectivas de análise, seus tipos e variantes.
- Casos particulares.
- Configurações de esquemas dimensionais.
- Algumas notas finais.



1

Introdução



Introdução

- No quotidiano de uma qualquer organização é usual tomarem-se **decisões**, qualquer que seja o seu ramo de atividade.
- Os intervenientes nesses processos - **agentes de decisão** - munem-se dos argumentos mais pertinentes para justificarem a sua opção por esta ou por aquela decisão e justificar, posteriormente, o seu bom ou mau resultado.
- A tomada de decisões não é um processo fácil.



Decisões e Decisões

- Existem **decisões** que conduzem a situações de bem-estar empresarial e “outras” cujo desfecho não é ambicionado por ninguém em situações ditas normais.
- Por vezes acredita-se que por trás de uma boa decisão está **uma boa “dose de sorte”**, algo que só o acaso conhece.
- Os agentes de decisão asseguram que para se tomar uma “boa” decisão é necessário um bom pacote de informação, selecionado de acordo com **as várias variáveis em jogo** no processo de decisão em causa.



Informação e Conhecimento

- **Informação e conhecimento** são um duo muito forte que, quando bem “conjugados”, tornam um processo de decisão simples e com resultados efetivos e mais agradáveis para a organização.
 - Qual o conhecimento que um agente de decisão deve possuir para realizar com sucesso as suas atividades mais mundanas de tomada de decisão?
 - Como é que a informação deve ser organizada de forma a constituir, de facto, um bem precioso no processo de tomada de decisão empresarial?



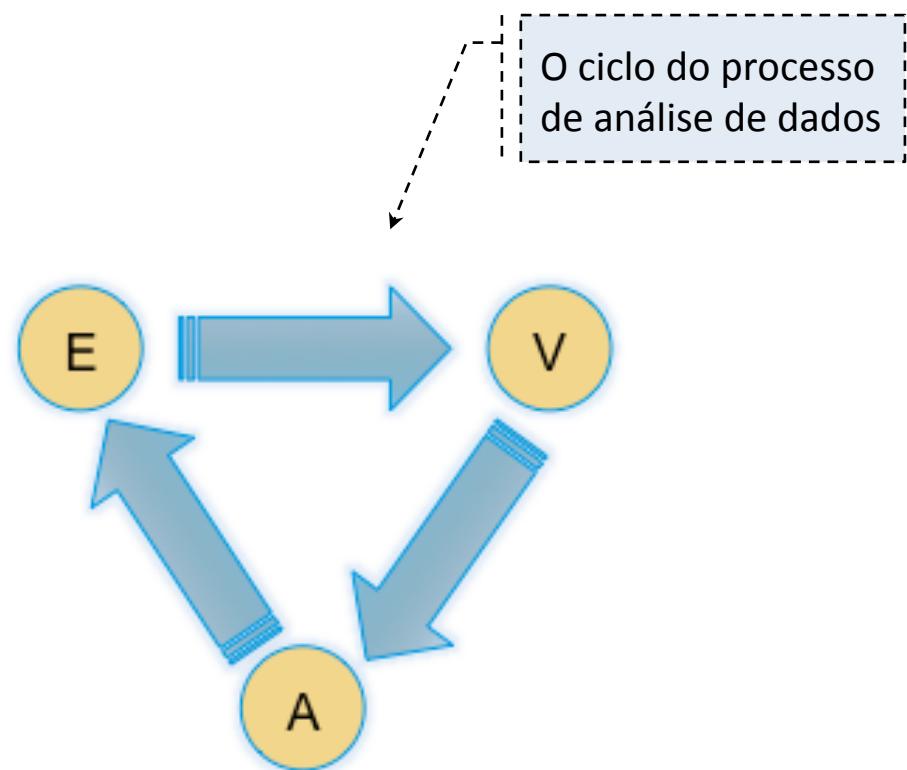
A Exploração da Informação

- Um agente de decisão não tem *a priori* uma agenda definida de *queries* ou de relatórios definidos.
- A forma como explora a informação segue a sua intuição, e a sua continuação, bem como a sua linha de investigação, obedece em grande medida aos resultados que uma (ou mais) das suas *queries* anteriores fez despoletar.



O Ciclo de Exploração

1. Extração de dados (frequentemente agregados).
2. Visualização de resultados (navegadores de dados ou *dashboards*).
3. Análise dos resultados.

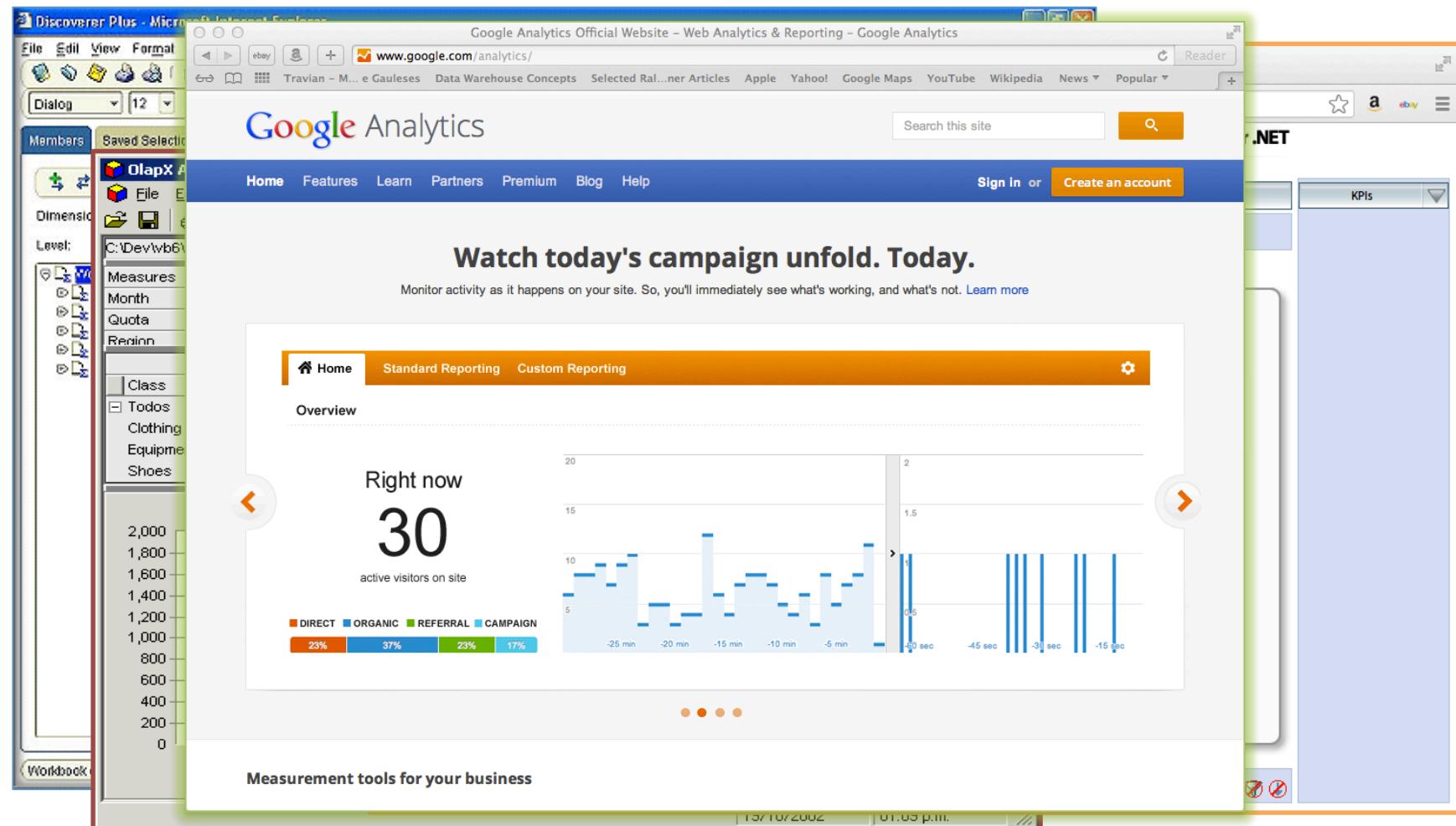


O Processamento Analítico

- É em OLAP (On-Line Analytical Processing) (Chaudhuri e Dayal, 1997) (Abello e Romero, 2009) que a excelência da exploração de estruturas de dados para a suporte à decisão aparece.
- As estruturas multidimensionais de dados (hipercubos) são as responsáveis pelo suporte de qualquer processo de exploração de dados e, como tal, vitais para o desempenho adequado de qualquer sistema orientado para a tomada de decisão.



O “Fim” a Alcançar



2

Desenvolvimento de Sistemas de *Data Warehousing*



Modelação Dimensional

- A **modelação dimensional de dados** (Kimbal e Ross, 2002) (Kimball, et al., 2008) é uma das atividades **mais relevantes** que usualmente se desenvolve no âmbito de um projeto de concepção e implementação de um sistema de *data warehousing* (SDW).

A modelação dimensional de dados é a atividade relacionada com o desenvolvimento de esquemas para sistemas de dados, especialmente orientados para o suporte a processos de tomada de decisão, cuja organização reflete a forma como os factos relacionados com uma ou mais vertentes de negócio podem ser explorados, de acordo com as várias perspetivas de análise de um ou mais agentes de decisão empresariais.



Modelos Dimensionais

- Os modelos dimensionais são os *alicerces* de todos os processos de tomada a decisão suportados por um *data warehouse*.
- A sua importância acentua-se se pensarmos que os esquemas dimensionais condicionam diretamente o próprio sistema de povoamento de um SDW.
 - Esquemas dimensionais complexos exigem sistemas de povoamento complexos.



Desenvolvimento de SDW

- Os SDW (Inmon, 1996) (Kimball, 1996) requerem **metodologias de concepção específicas**, adequadas à sua própria natureza e objetivos.
- Mesmo quando de inspiração operacional, os SDW devem acolher e refletir **as necessidade de suporte à decisão** requeridas pelos agentes empresariais e serem capazes de satisfazer as suas *queries* e necessidades de *reporting*.

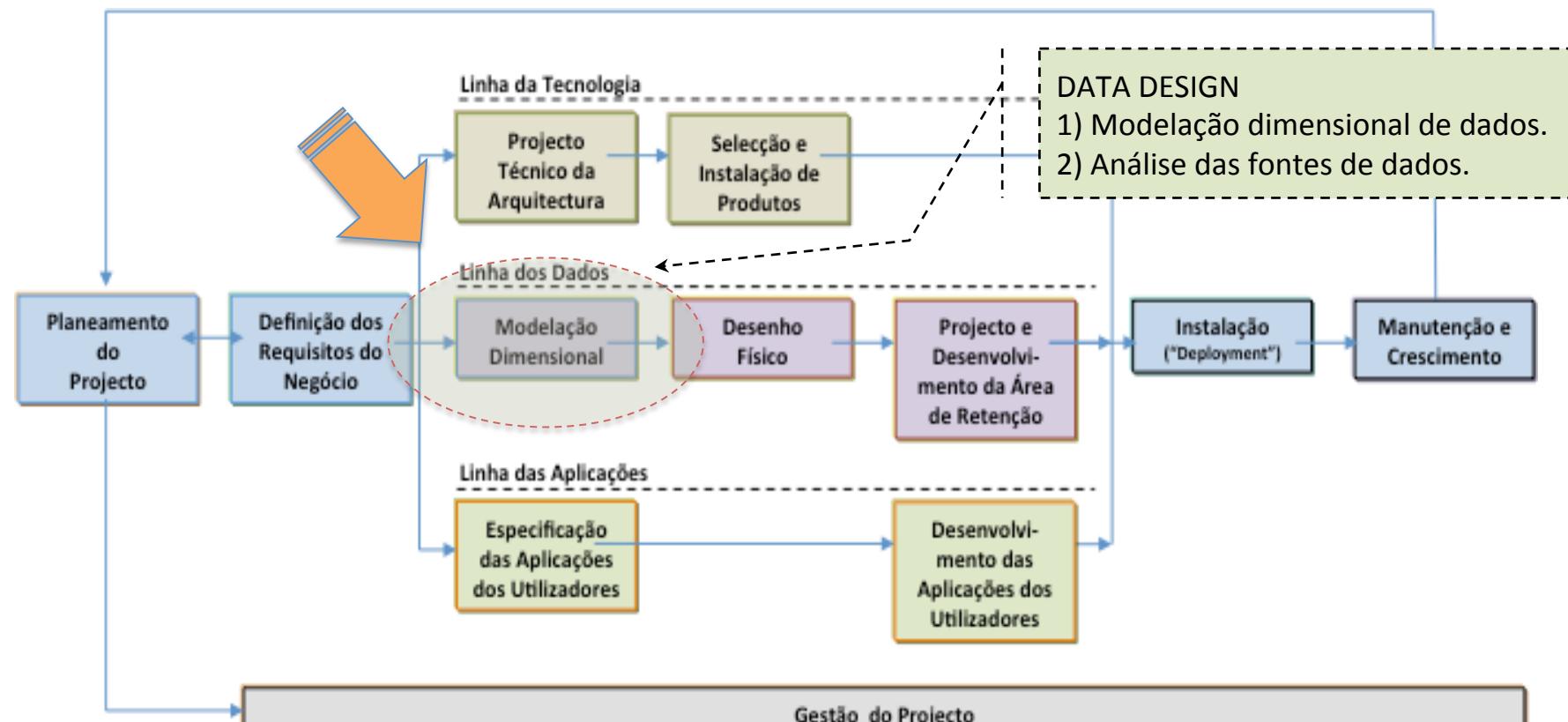


O Processo de Desenvolvimento

- Os SDW devem também ser **rápidos**, apresentando excelentes desempenhos na satisfação desses requisitos, o que implica que as suas **estruturas de dados** devam estar **orientadas especificamente para o fornecimento dos dados** pedidos sem necessidade de:
 - *queries complexas que imponham processos exigentes de combinação de dados contidos em várias tabelas*



O Ciclo de Desenvolvimento



(Kimball, et al., 1998)



A Etapa da Modelação Dimensional

1. Construção da matriz de decisão.
2. Seleção do *data mart* a desenvolver.
3. Escolha do grão das tabelas de factos.
4. Escolha das dimensões de análise.
5. Desenvolver o diagrama das tabelas de factos.
6. Documentar as tabelas de factos.
7. Projetar o detalhe das dimensões.
8. Desenvolver os diversos factos derivados.
9. Revisão do projeto com os utilizadores e sua aceitação.
10. Revisão das recomendações de ferramentas end-user para o projeto da base de dados.



A Etapa da Modelação Dimensional

11. Revisão das recomendações de sistemas de gestão de bases de dados para o projeto da base de dados.
12. Completar o esquema lógico da base de dados.
13. Identificar os possíveis candidatos de agregados armazenados previamente.
14. Desenvolver a estratégia de desenvolvimento para as tabelas de agregados.
15. Revisão do esquema lógico da base de dados.
16. Certificar o esquema desenvolvido para a base de dados com o fornecedor das ferramentas para suporte à decisão.
17. Rever o projeto e tratar da aceitação por parte dos utilizadores.



A Etapa da Análise das Fontes

1. Identificar as fontes de dados candidatas.
2. Analisar o conteúdo das fontes de dados – dados e metadados.
3. Desenvolver uma tabela com o mapeamento dos dados entre as diversas fontes de dados operacionais e os dados do *data warehouse* – *source-to-target map*.
4. Estimar o número de registos envolvidos futuramente no processo de povoamento.
5. Rever o projeto e tratar da aceitação por parte dos seus futuros utilizadores.



Notações e Modelos
Metodologias de Construção

3

Construção de *Data Warehouses* Metodologias e Esquemas Dimensionais



Notações e Modelos

- A **modelação de esquemas** para sistemas de dados sempre foi uma área que despertou muito interesse.
- A forma como se capta e representa **os requisitos dos utilizadores** de um sistema de dados é um fator determinante, não só na própria **representação** do sistema como, mais tarde, na sua **operacionalidade e qualidade de serviço**.
- Muitos modelos e notações têm sido apresentadas ao longo dos tempos, abordando (e justificando) de **diferentes maneiras diferentes abordagens** do ponto de vista sintático e semântico.



Notações e Modelos

- Desde **Chen (1976)** que temos assistido à emergência de inúmeras notações para a representação conceptual de sistemas de dados - **Barker, IDEF1X, ORM, UML ou XML**.
- Nem sempre é fácil argumentar **a favor ou contra esta ou aquela notação**, pelo simples motivo que em certos aspectos cada uma delas apresenta **visões diferentes** para a representação e tratamento dos dados.
- Mas, apesar disso, não será difícil observar a relevância das notações mais recentes, como **a ORM** (Halpin & Morgan, 2008), pelo seu poder de **expressão sintático e semântico**.

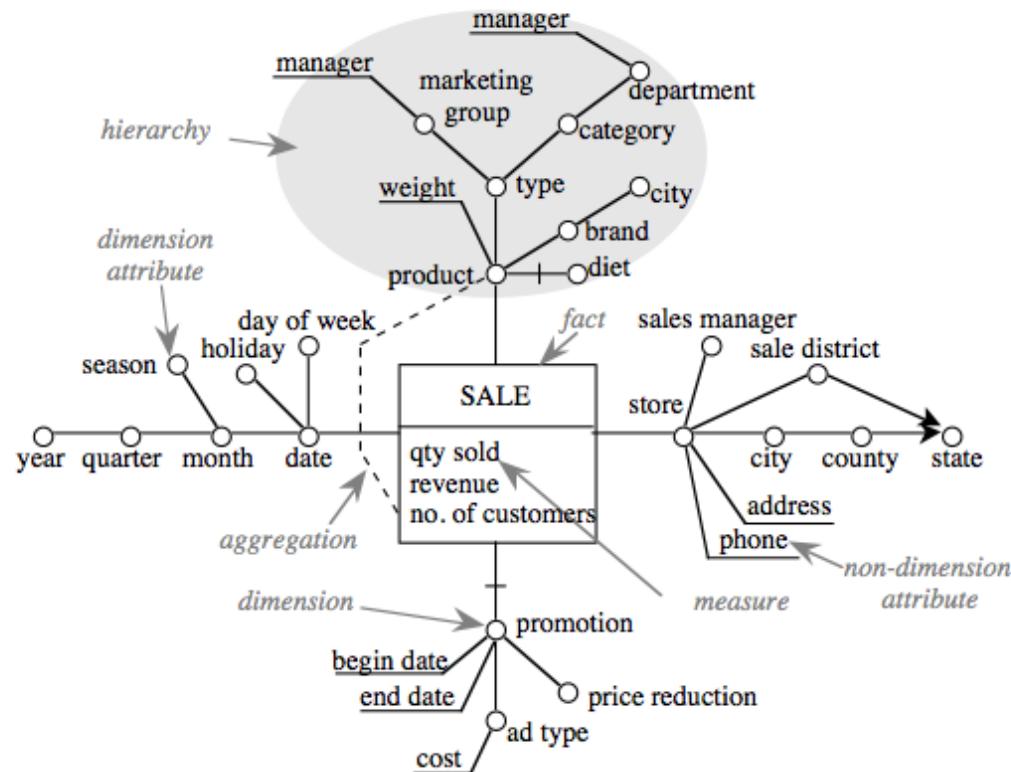


Uma Notação

- Uma das primeiras notações propostas exclusivamente para a modelação conceptual de um esquema dimensional para um *data warehouse* foi proposta por Golfarelli, et al. (1998):
 - *Dimensional Fact model* (DFM).
- O conjunto dos elementos básicos da notação de Golfarelli, et al. (1998) inclui representações para:
 - factos, atributos, dimensões e hierarquias;
 - medidas de um facto, sendo ou não ser aditiva (agregável); em que estes elementos são conjugados através de esquemas de factos organizados numa estrutura em árvore.



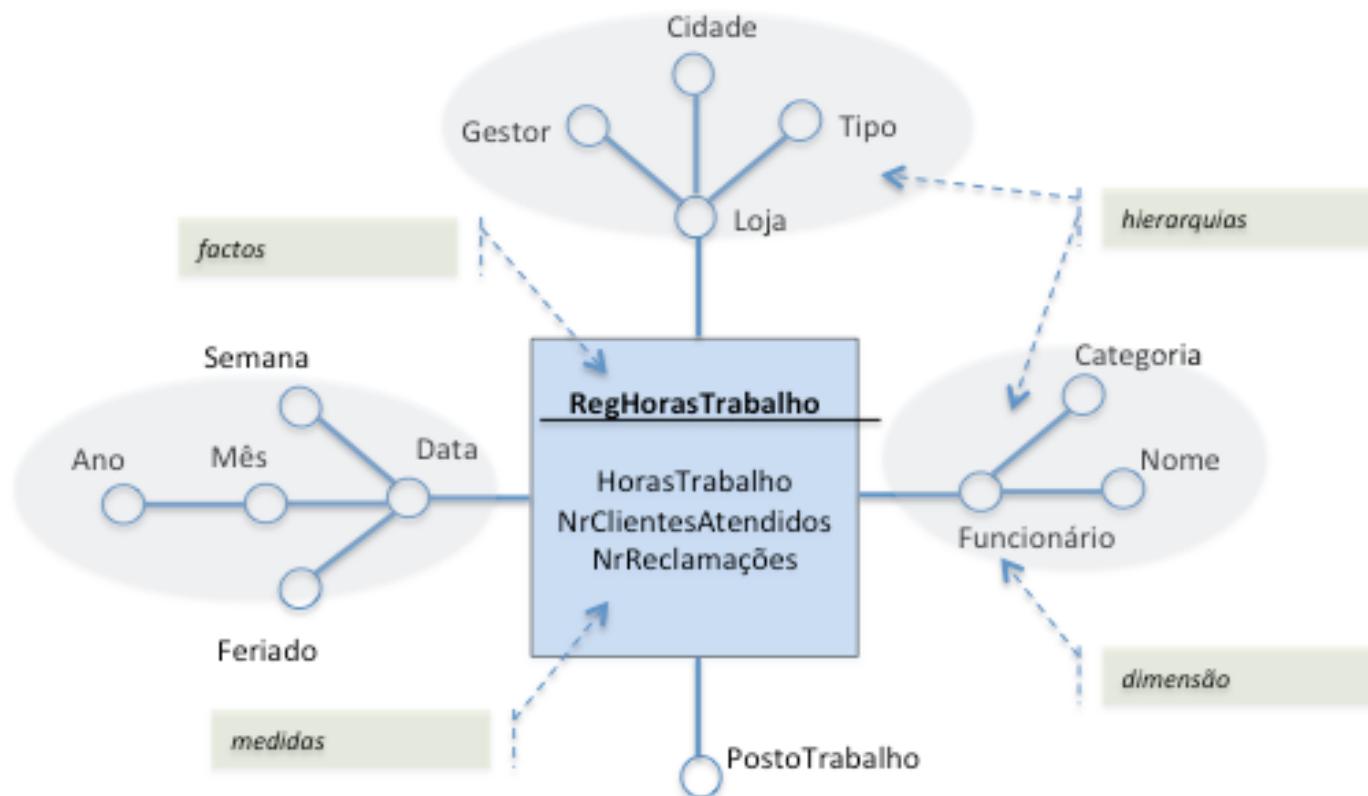
A Notação de Golfarelli, et al. (1998)



Fonte: M. Golfarelli, D. Maio, S. Rizzi. The Dimensional Fact Model: a Conceptual Model for Data Warehouses. Invited paper. International Journal of Cooperative Information Systems, vol. 7, n. 2&3, 1998.



A Notação de Golfarelli, et al. (1998)



Metodologias de Construção

- Hüsemann, et al. (2000) - o desenho conceptual deve ser claramente independente de qualquer sistemas de gestão de bases de dados alvo.
- Cabibbo e Torlone (2000) - acrescentam um nível de abstração lógico para garantir a independência entre as aplicações OLAP e a estrutura física do *data warehouse*.
- Vassiliadis (2000) - um hipercubo não é um entidade por si, mas sim uma vista sobre um dado conjunto de dados.
- Tsois, et al. (2001) - a modelação dimensional do ponto de vista dos requisitos do utilizador final de aplicações OLAP reais.
- Luján-Mora e Trujillo (2003) - um método geral e standard para o desenho de data warehouses assente na UML.
- Jones e Song (2005) - uma metodologia que assenta na utilização de patterns – dimensional design patterns (DDP) – e suas aplicações.
- Dori, et al. (2008) - o método *Object-process-based Data Warehouse Construction*.
- Romero & Albello (2009) - uma técnica centrada no utilizador final (user-centered) para fazer a validação dos seus requisitos e das tarefas de desenho multidimensional.
- (...)



4

A Adoção de um Modelo



O Processo de Modelação

- O processo de modelação dimensional que iremos seguir acompanhará de perto **a abordagem proposta por Kimball e Ross (2002)**, desenvolvendo-o, passo a passo, requisito a requisito, de forma a desenvolver um esquema multidimensional cobrindo todos os tipos de objetos de dados – **tabelas de facto, dimensões, tabelas ponte e medidas** – que podemos encontrar neste tipo de esquemas.

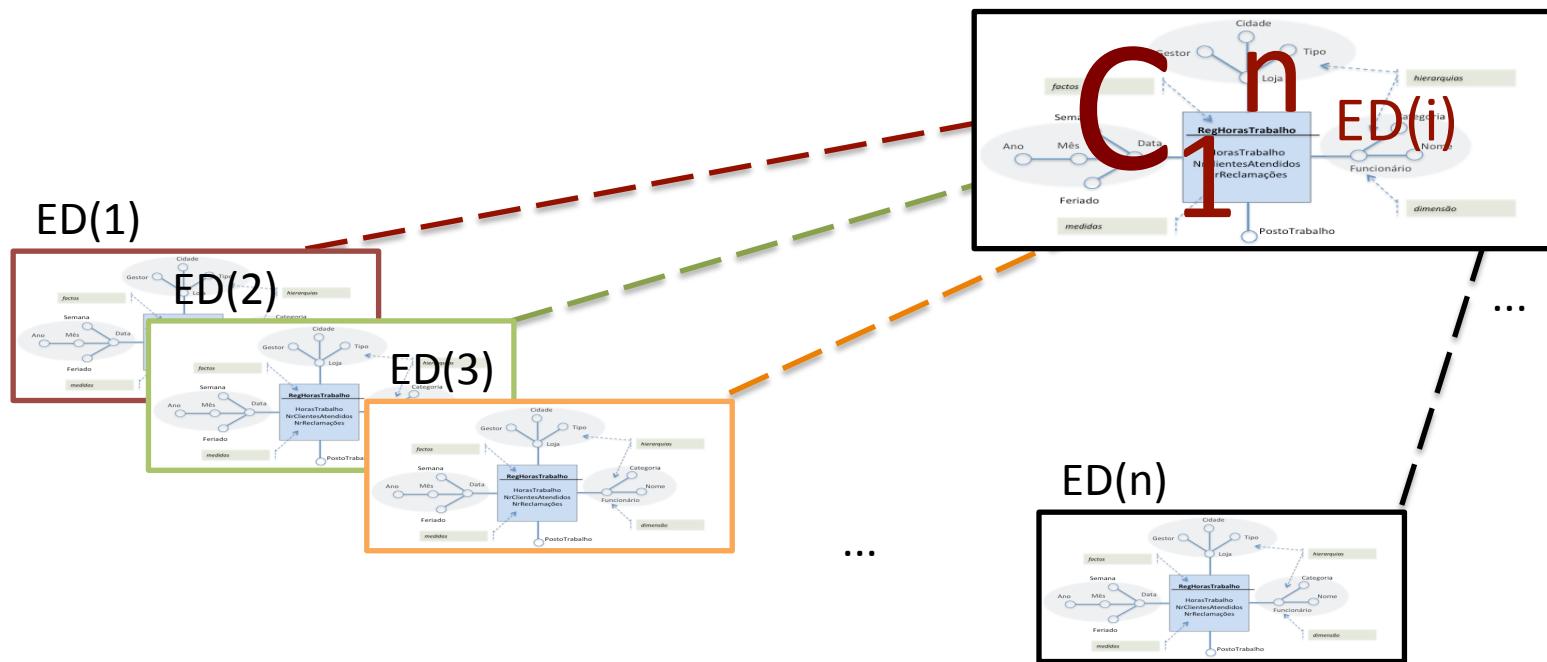


Um Desenvolvimento *Bottom-Up*

- Uma das formas mais usuais de fazer o desenvolvimento de um esquema dimensional é através da utilização do método dos “4 passos” (Kimball e Ross 2002), que pressupõe o desenvolvimento do sistema de data warehousing tipicamente **de baixo para cima (bottom-up)**.
- O “nosso” data warehouse será projetado **área a área**, sendo desenvolvido **de forma incremental** e tomando em consideração todos os (sub)esquemas dimensionais desenvolvidos até ao momento.



Bottom-Up na Prática



O Método do “4 Passos”

- Os quatro passos do método são (Kimball e Ross, 2002) (Imhoff, et al, 2003):
 1. Seleção da área de suporte à decisão a implementar.
 2. Definição do detalhe dos factos (**o grão**) do processo selecionado.
 3. Seleção das dimensões de análise sobre as quais se pretende analisar os factos.
 4. Definição das medidas a integrar na estrutura de cada facto.



O Contexto de Decisão
O Caso de Estudo
Identificação dos Processos de Negócio

5

Seleção da Área de Suporte à Decisão



A Área de Negócio

- Na modelação dimensional de um *data warehouse* devemos preservar a imagem global da organização, de forma a não perder **o horizonte dos processos de tomada de decisão como um todo**.
- O processo de modelação dimensional inicia-se com a **identificação e caracterização da área de negócio** em que desejamos desenvolver as nossas atividades de tomada de decisão - Definição do projeto e Gestão e planeamento do projeto (Kimball, et al., 1998).



O Problema

A “L&LNet” é uma empresa que detém uma rede de livrarias espalhadas por grande parte do continente europeu. A sua atividade comercial desenvolve-se em todos os domínios do conhecimento, vendendo, promovendo ou simplesmente divulgando livros de todos os géneros literários, escritos nas mais diversas línguas. Como resultado de uma política concorrencial aguerrida, a “L&LNet” cresceu nos últimos 5 anos cerca de 25% em termos de número de lojas abertas ao público e cerca de 42.5% no seu volume de vendas total. O sucesso da sua atividade deve-se em grande parte ao seu quadro de gestores internacionais, que se preocupam permanentemente em munir as lojas de cada cidade, de cada país, com os livros que a sua população tem tendência para comprar. Para isso, realizam regularmente, pelo menos uma vez de três em três meses, um trabalho exaustivo de pesquisa (com a aplicação de inquéritos pessoais aos seus clientes) para identificação e apreciação das suas tendências de leitura. Na realidade, a “LNet” não faz mais do que uma simples operação de profiling dos seus clientes. Assim, consegue apresentar aos seus clientes nas suas lojas as últimas novidades editoriais, bem como sugerir “pacotes” de leitura a preços promocionais.

(...)



O Porquê deste Problema

- É um caso de estudo puramente **fictício**.
- Dá-nos um **grande grau de liberdade** no processo de modelação – é um problema bastante genérico.
- Não nos vincula a estruturas de dados específicas.
- Permite-nos desenvolver e apresentar um modelo de dados dimensional ao “sabor” da **exposição** de cada tópico.



Alguns Motivos e Justificações

- Desenvolver ações de incentivo comercial, à medida.
- Incentivar as vendas de livros.
- Estabelecer rankings de clientes.
- Reduzir os livros em stock.
- Definir um ranking de funcionários.
- Melhorar a qualidade de serviço de vendas.
- (...)



A Matriz de Decisão

Caracterização de Data Mart Comercial																																																		
Identificação: Comercial																																																		
Descrição Geral: Informação para suporte à tomada de decisão na área de vendas da "L&LN" providenciando elementos de dados seleccionados acerca das vendas de livros em todas as suas lojas, para gestão e controlo das ações comerciais realizadas e dos stocks envolvidos e para fazer a avaliação do desempenho dos seus funcionários ao longo de um dia de trabalho, com base nos pontos que lhes vão sendo atribuídos a partir das vendas concretizadas.																																																		
Estrutura base																																																		
<table border="1"><thead><tr><th>Tabelas de Factos >></th><th>TF-Vendas</th><th>TF-Pontos</th></tr></thead><tbody><tr><td><< Dimensões</td><td></td><td></td></tr><tr><td>Calendário</td><td>✓</td><td>✓</td></tr><tr><td>Lojas</td><td>✓</td><td>✓</td></tr><tr><td>Clientes</td><td>✓</td><td></td></tr><tr><td>Funcionários</td><td>✓</td><td>✓</td></tr><tr><td>Livros</td><td>✓</td><td></td></tr><tr><td>Editoras</td><td>✓</td><td></td></tr><tr><td>Autores</td><td>✓</td><td></td></tr><tr><td>Géneros</td><td>✓</td><td></td></tr><tr><td>Línguas</td><td>✓</td><td></td></tr><tr><td>Vendas</td><td>✓</td><td></td></tr><tr><td>Pontos de Venda</td><td>✓</td><td></td></tr><tr><td>Países</td><td>✓</td><td></td></tr><tr><td>Períodos de Trabalho</td><td></td><td>✓</td></tr><tr><td>Clima</td><td>✓</td><td>✓</td></tr></tbody></table>			Tabelas de Factos >>	TF-Vendas	TF-Pontos	<< Dimensões			Calendário	✓	✓	Lojas	✓	✓	Clientes	✓		Funcionários	✓	✓	Livros	✓		Editoras	✓		Autores	✓		Géneros	✓		Línguas	✓		Vendas	✓		Pontos de Venda	✓		Países	✓		Períodos de Trabalho		✓	Clima	✓	✓
Tabelas de Factos >>	TF-Vendas	TF-Pontos																																																
<< Dimensões																																																		
Calendário	✓	✓																																																
Lojas	✓	✓																																																
Clientes	✓																																																	
Funcionários	✓	✓																																																
Livros	✓																																																	
Editoras	✓																																																	
Autores	✓																																																	
Géneros	✓																																																	
Línguas	✓																																																	
Vendas	✓																																																	
Pontos de Venda	✓																																																	
Países	✓																																																	
Períodos de Trabalho		✓																																																
Clima	✓	✓																																																
<table border="1"><thead><tr><th>Número Dimensões</th><th>13</th><th>4</th></tr></thead><tbody><tr><td>Tipo</td><td>Transacional</td><td>Instantânea</td></tr><tr><td>Periodicidade</td><td>Diária</td><td>Diária</td></tr><tr><td>Descrição</td><td>Transações comerciais de livros.</td><td>Atribuição de pontos de desempenho.</td></tr><tr><td>Utilidade estratégica</td><td>Avaliação do desempenho comercial de cada uma das lojas. Incentivar as vendas de livros. Identificar e caracterizar nichos de mercado. Definição de ações promocionais. Estabelecer um ranking de clientes. Definição e caracterizar de perfis de vendas de livros. Otimização de stocks. Melhorar base de negociação com fornecedores.</td><td>Ranking de funcionários. Atribuição de prémios de produtividade. Definição de perfis profissionais.</td></tr><tr><td>Utilizadores</td><td>Administradores gerais e gestores de loja.</td><td>Administradores gerais, gestores de loja e chefes de pessoal.</td></tr><tr><td>Observações</td><td colspan="2" rowspan="2">Nada a assinalar.</td></tr></tbody></table>			Número Dimensões	13	4	Tipo	Transacional	Instantânea	Periodicidade	Diária	Diária	Descrição	Transações comerciais de livros.	Atribuição de pontos de desempenho.	Utilidade estratégica	Avaliação do desempenho comercial de cada uma das lojas. Incentivar as vendas de livros. Identificar e caracterizar nichos de mercado. Definição de ações promocionais. Estabelecer um ranking de clientes. Definição e caracterizar de perfis de vendas de livros. Otimização de stocks. Melhorar base de negociação com fornecedores.	Ranking de funcionários. Atribuição de prémios de produtividade. Definição de perfis profissionais.	Utilizadores	Administradores gerais e gestores de loja.	Administradores gerais, gestores de loja e chefes de pessoal.	Observações	Nada a assinalar.																												
Número Dimensões	13	4																																																
Tipo	Transacional	Instantânea																																																
Periodicidade	Diária	Diária																																																
Descrição	Transações comerciais de livros.	Atribuição de pontos de desempenho.																																																
Utilidade estratégica	Avaliação do desempenho comercial de cada uma das lojas. Incentivar as vendas de livros. Identificar e caracterizar nichos de mercado. Definição de ações promocionais. Estabelecer um ranking de clientes. Definição e caracterizar de perfis de vendas de livros. Otimização de stocks. Melhorar base de negociação com fornecedores.	Ranking de funcionários. Atribuição de prémios de produtividade. Definição de perfis profissionais.																																																
Utilizadores	Administradores gerais e gestores de loja.	Administradores gerais, gestores de loja e chefes de pessoal.																																																
Observações	Nada a assinalar.																																																	
Versão 1.00/2012, Belo, O.																																																		

- Caracterização geral
- Estrutura base
- Tipo
- Periodicidade
- Utilidade
- Perfis de utilização



A Matriz de Decisão

<p><i>Caraterização de Data Mart Comercial</i></p>
<p>Identificação: Comercial</p>
<p>Descrição Geral: <i>Informação para suporte à tomada de decisão na área de vendas da "L&LNet" providenciando elementos de dados selecionados acerca das vendas de livros em todas as suas lojas, para gestão e controlo das ações comerciais realizadas e dos stocks envolvidos e para fazer a avaliação do desempenho dos seus funcionários ao longo de um dia de trabalho, com base nos pontos que lhes vão sendo atribuídos a partir das vendas concretizadas.</i></p>



A Matriz de Decisão

Estrutura base		
Tabelas de Factos >>	TF-Vendas	TF-Pontos
<< Dimensões		
Calendário	✓	✓
Lojas	✓	✓
Clientes	✓	
Funcionários	✓	✓
Livros	✓	
Editoras	✓	
Autores	✓	
Géneros	✓	
Línguas	✓	
Vendas	✓	
Pontos de Venda	✓	
Países	✓	
Períodos de Trabalho		✓
Clima	✓	✓



A Matriz de Decisão

Número Dimensões	13	4
Tipo	Transacional	Instantânea
Periodicidade	Diária	Diária
Descrição	Transações comerciais de livros.	Atribuição de pontos de desempenho.
Utilidade estratégica	Avaliação do desempenho comercial de cada uma das lojas. Incentivar as vendas de livros. Identificar e caracterizar nichos de mercado. Definição de ações promocionais. Estabelecer um ranking de clientes. Definição e caracterizar de perfis de vendas de livros. Otimização de stocks. Melhorar base de negociação com fornecedores.	Ranking de funcionários. Atribuição de prémios de produtividade. Definição de perfis profissionais.
Utilizadores	Administradores gerais e gestores de loja.	Administradores gerais, gestores de loja e chefes de pessoal.
Observações	Nada a assinalar.	
	Versão 1.00/2012, Belo, O.	



As Dimensões do DM Comercial

Dimensões do Data Mart "Comercial"			
Nº	Identificação	Descrição	Esquema (Tipo)
1	Calendário	Esta é a dimensão temporal. Acolhe todos os atributos que sustentem análises ao longo do tempo, como data, mês, semana, feriado, etc.	Dim-Calendário (Com diferentes papéis).
2	Lojas	Caraterização das lojas que integram a rede da "L&LNet".	Dim-Loja (Com variação).
3	Clientes	Identificação e caraterização dos clientes das diversas lojas.	Dim-Cliente (Com variação), Dim-Cliente-HST (Histórico) e Mini-Dim-Cliente (Mini dimensão).
4	Funcionários	Identificação e caraterização dos clientes de cada uma das lojas.	Dim-Funcionário (Com variação) e AT-LínguasFuncionários (Ponte).
5	Livros	Informação sobre o catálogo geral dos livros à venda nas lojas, complementada com informação relativa a stocks e vendas passadas.	Dim-Livro (Com variação).
6	Editoras	Dados gerais sobre as editoras dos livros.	Dim-Editora (Normal).
7	Autores	Identificação e caraterização dos autores dos livros.	Dim-Autor (Com variação), Dim-AutorGrupo (Ponte), BT-Autor (Ponte).
8	Géneros	Géneros literários dos livros à venda nas lojas.	Dim-Género (Normal)
9	Línguas	Línguas em que estão escritos os livros.	Dim-Língua (Normal)
10	Vendas	Número dos documentos de venda.	Dim-Controloloja (De controlo)
11	Pontos de Venda	Números dos ponto de venda das diversas lojas.	
12	Países	Identificação e caraterização dos países nos quais a "L&LNet" tem lojas.	Dim-País (Normal)
13	Periodos de Trabalho	Caraterização dos diversos períodos de trabalho ao longo de um dia de vendas.	Dim-Periodo-Trabalho (Regular)
14	Clima	Informação sobre as condições climatéricas que se verificavam nos dias das vendas de livros.	Dim-Clima (Normal)
15	Localidade	Dados sobre as localidades de residência dos clientes.	Dim-Localidade (Subdimensão)
16	Zona	Dados sobre as zonas de residência dos clientes.	Dim-Zona (Subdimensão)

A lista completa das diversas dimensões (e objetos de dados relacionados) que irão integrar o data mart "Comercial" da "L&LNet".



As Dimensões do DM Comercial

Nº	Identificação	Descrição	Esquema (Tipo)
1	Calendário	Esta é a dimensão temporal. Acolhe todos os atributos que sustentem análises ao longo do tempo, como data, mês, semana, feriado, etc.	Dim-Calendário (Com diferentes papéis).
2	Lojas	Caraterização das lojas que integram a rede da "L&LNet".	Dim-Loja (Com variação).
3	Clientes	Identificação e caraterização dos clientes das diversas lojas.	Dim-Cliente (Com variação), Dim-Cliente-HST (Histórico) e Mini-Dim-Cliente (Mini dimensão).
4	Funcionários	Identificação e caraterização dos clientes de cada uma das lojas.	Dim-Funcionário (Com variação) e AT-LínguasFuncionários (Ponte).
5	Livros	Informação sobre o catálogo geral dos livros à venda nas lojas, complementada com informação relativa a stocks e vendas passadas.	Dim-Livro (Com variação).



Grão de uma Tabela de Factos
As Tabelas de Factos

6

Grão, Factos e sua Representação



O Grão de uma Tabela de Factos

- Alcançámos uma das etapas mais delicadas da modelação dimensional.
- A definição do detalhe da informação que queremos manter nas nossas estruturas de dados do data warehouse - **o grão da tabela de factos**.
- A definição do grão está intimamente ligada com a estrutura da tabela de factos e é uma das tarefas mais difíceis de se realizar.



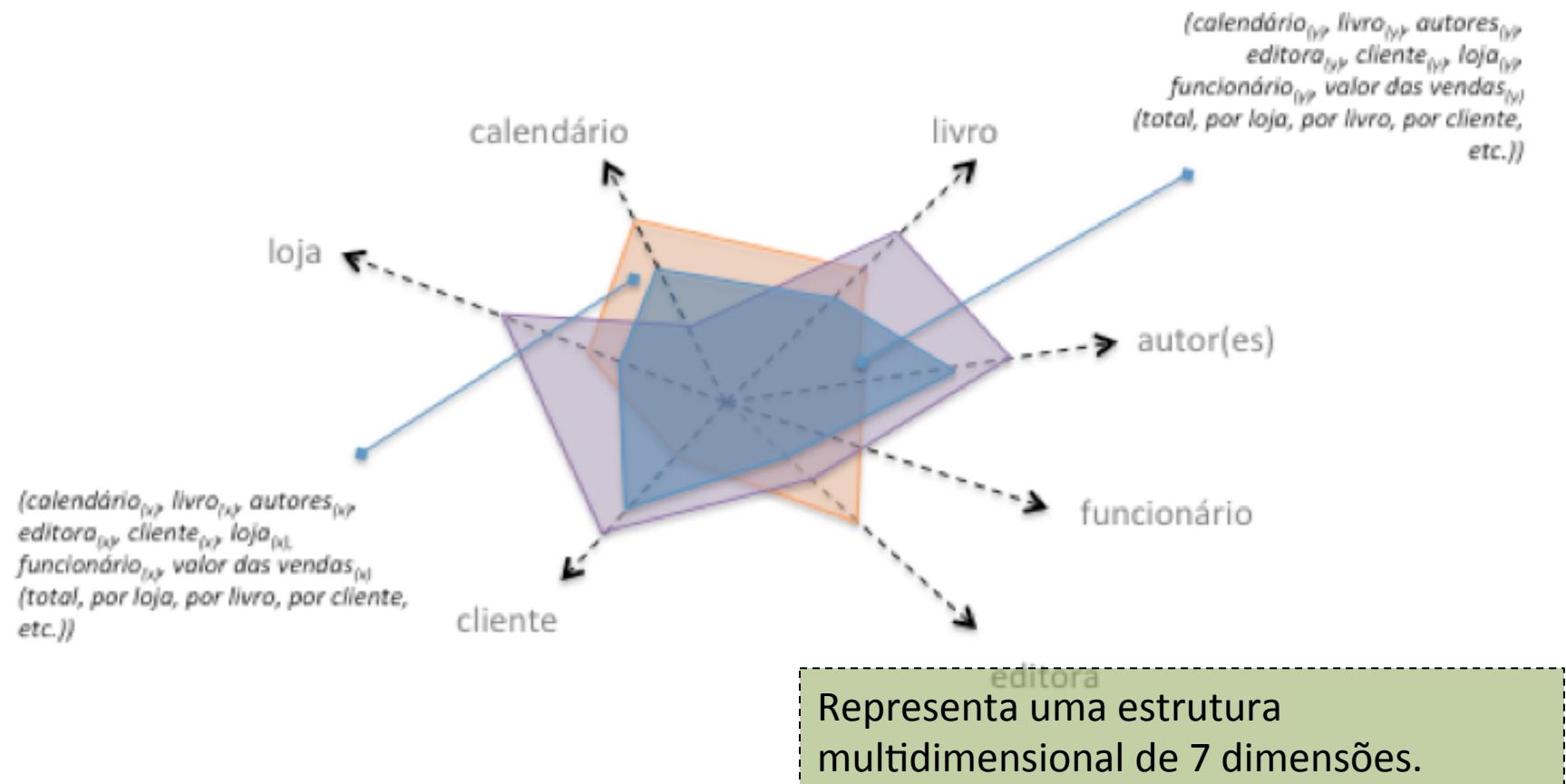
Um Cenário Típico

Para suportar os processos de decisão direta ou indiretamente ligados com as vendas de livros, os agentes de decisão da “L&LNet” precisam de saber, simplesmente, quais os livros que foram vendidos em cada uma das suas lojas, quando e a quem. A partir dessa informação, conseguir-se-á saber praticamente tudo sobre as coisas mais elementares relacionadas com uma venda: valor das vendas (total, por loja, por livro, por cliente, etc.). Apesar disso, querem que essa informação seja enriquecida com outros dados, para que possam também aferir a influência (ou a importância) da editora (fornecedor), dos seus autores e, por fim, dos funcionários da própria loja.

(...)



Uma Star-Net Query



Complementando o Cenário

(...)

Como queremos analisar, também, as vendas dos livros em termos internacionais – integramos os dados de todas as lojas distribuídas pelo mundo -, é muito útil poder identificar as tendências dos vários clientes em termos de géneros literários e línguas de leitura escolhidas, em cada país em que temos uma loja.

(...)

Este “complemento” acrescenta três novos eixos de análise, três novas dimensões: género literário do livro, língua em que o livro foi escrito e o país no qual foi efetuada a venda do livro.



A Definição de Grão

- O grão de uma tabela de factos corresponde define a estrutura base, mais refinada, dos seus registos.
 - Não existe mais detalhe para além do grão.
 - O grão só pode ser agregado.
 - Não há maneira de desagregar o grão.
 - O grão é aquele que define o nível de informação mais atómico.
- Uma má definição do grão (fará com que a exploração de dados de uma tabela de factos dê origem a resultados inconsistentes ou pouco coerentes – não se devem “**misturar alhos com bugalhos**”.



O Grão

Versão 1

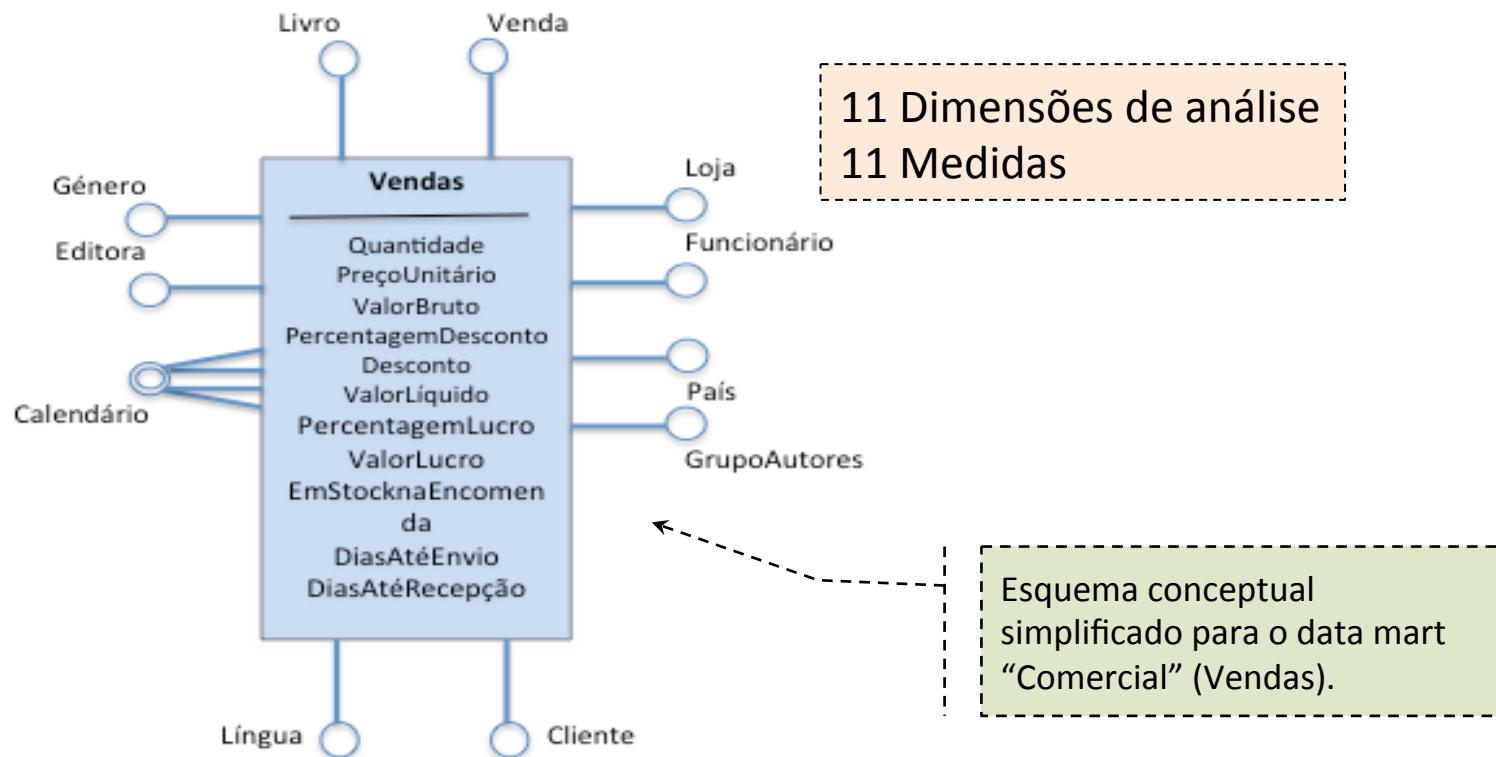
"A venda de um ou mais exemplares de um livro, a um cliente específico, efectuada numa loja da "L&LNet", atendida por um funcionário, num dado dia (calendário)."

Versão 2

"A venda de um ou mais exemplares de um livro (de um dado género literário, escrito numa dada língua por um conjunto de autores específico, para uma editora em particular), a um cliente específico, efectuada numa loja da "L&LNet", atendida por um funcionário, num dado dia (calendário), num certo país."



Um primeiro Esquema



Uma Tabela de Factos

Caracterização de Tabela de Factos												
Identificação	TF-Vendas.											
Descrição	Tabela que acolhe todos os registos de vendas de livros realizados nas várias lojas da "L&LNet".											
Data Mart	Comercial.											
Tipo	Transacional.											
Utilidade estratégica	Incentivar as vendas de livros. Estabelecer um ranking de clientes para lançamento de ações promocionais. Identificar e caracterizar as categorias dos livros vendidos para otimizar stocks.											
Povoamento	Realizado diariamente, entre a uma e as sete horas da manhã, iniciando-se, de preferência a sua execução às duas da manhã.											
Dimensão Inicial	1KR (234KB), após primeiro povoamento.											
Crescimento	10%/mês.											
Período de dados	Os 4 últimos anos de vendas de livros. Os dados de anos anteriores ficarão em arquivo.											
Atributos												
Dimensões												
Nº	Identificação	Chave	Tipo	Domínio	Descrição	Exemplo						
1	CódigoVendaId	S	J	Inteiro	Código interno do documento de venda emitido pela loja ao cliente.	8732651						
2	LoyaltyId	S	V	Inteiro	Código interno para identificação da loja da "L&LNet" na qual a venda foi realizada.	1						
3	FuncionarioId	S	V	String(1)	Código interno para a identificação do período de trabalho de cada loja.	M						
4	PeriodoTrabalhoId	S	N	Inteiro	Código interno para a identificação do funcionário que atendeu o cliente e processou a venda.	6						
5	DtEncomenda	S	RP	Data	Data em que o livro foi encomendado. Numa venda efetuada diretamente na loja esta data corresponde à data da concretização da venda.	2011/12/05						
6	DtVenda	S	RP	Data	Data em que a venda foi realizada.	2011/12/05						
7	DtEnvio	S	RP	Data	Data em que o livro foi enviado para o cliente. Numa venda efetuada diretamente na loja esta data corresponde à data da concretização da venda. Numa venda efetuada diretamente na loja esta data corresponde à data da concretização da venda.	2011/12/05						
8	DtRecepção	S	RP	Data	Data em que o livro foi recepcionado pelo cliente. Numa venda efetuada diretamente na loja esta data corresponde à data da concretização da venda.	2011/12/05						
9	PaísId	S	N	Inteiro	Código interno para a identificação do país em que a venda do livro foi realizada.	1						
10	ClienteId	S	V	Inteiro	Código interno para o cliente da "L&LNet". Caso a venda não tenha sido associada com um cliente cadastrado, este código terá o valor '0'.	0						
11	GrupoAutoresId	S	N-BT	Inteiro	Código interno para a identificação do grupo de autores que escreveu o livro.	1						
12	LínguaId	S	N	Inteiro	Código interno para a língua em que o livro foi escrito.	1						
13	GêneroId	S	N	Inteiro	Código interno que identifica o género literário de um livro.	1						

- Caracterização geral.
- Atributos de dimensão e medidas.
- Índices.
- Perfis de utilização.



Uma Tabela de Factos

Caracterização de Tabela de Factos	
Identificação	TF-Vendas.
Descrição	Tabela que acolhe todos os registos de vendas de livros realizados nas várias lojas da "L&LNet".
Data mart	Comercial.
Tipo	Transacional.
Utilidade estratégica	Incentivar as vendas de livros. Estabelecer um ranking de clientes para lançamento de ações promocionais. Identificar e caracterizar as categorias dos livros vendidos para otimizar stocks.
Povoamento	Realizado diariamente, entre a uma e as sete horas da manhã, iniciando-se, de preferência a sua execução às duas da manhã.
Dimensão inicial	1KR (234KB), após primeiro povoamento.
Crescimento	10%/mês.
Período de dados	Os 4 últimos anos de vendas de livros. Os dados de anos anteriores ficarão em arquivo.



Uma Tabela de Factos

Atributos						
Dimensões						
Nr	Identificação	Chave	Tipo	Domínio	Descrição	Exemplos
1	<u>CódigoVendaId</u>	S	J	Inteiro	Código interno do documento de venda emitido pela loja ao cliente	8732651
2	<u>LojaId</u>	S	V	Inteiro	Código interno para identificação da loja da "L&LNet" na qual a venda foi realizada.	1
3	<u>FuncionárioId</u>	S	V	String (1)	Código interno para a identificação do período de trabalho de cada loja.	M
4	<u>PeríodoTrabalhoId</u>	S	N	Inteiro	Código interno para a identificação do funcionário que atendeu o cliente e processou a venda.	6
5	<u>DtEncomenda</u>	S	RP	Data	Data em que o livro foi encomendado. Numa venda efetuada diretamente na loja esta data corresponde à data da concretização da venda.	2011/12/05
6	<u>DtVenda</u>	S	RP	Data	Data em que a venda foi realizada.	2011/12/05
7	<u>DtEnvio</u>	S	RP	Data	Data em que o livro foi enviado para o cliente. Numa venda efetuada diretamente na loja esta data corresponde à data da concretização da venda. Numa venda efetuada diretamente na loja esta data corresponde à data da concretização da venda.	2011/12/05
8	<u>DtRecepção</u>	S	RP	Data	Data em que o livro foi recepcionado pelo cliente. Numa venda efetuada diretamente na loja esta data corresponde à data da concretização da venda.	2011/12/05
9	<u>PaísId</u>	S	N	Inteiro	Código interno para a identificação do país em que a venda do livro foi realizada.	1
10	<u>ClienteId</u>	S	V	Inteiro	Código interno para o cliente da "L&LNet". Caso a venda não tenha sido associada com um cliente cadastrado, este código terá o valor '0'.	0
11	<u>GrupoAutoresId</u>	S	N-BT	Inteiro	Código interno para a identificação do grupo de autores que escreveu o livro.	1
12	<u>LínguaId</u>	S	N	Inteiro	Código interno para a língua em que o livro foi escrito.	1



Uma Tabela de Factos

Medidas					
Nr	Identificação	Domínio	Tipo (Função)	Descrição	Exemplos
1	Quantidade	Inteiro	A (sum)	Número de exemplares vendidos do livro.	2
2	PreçoUnitário	Decimal(19,2)	N	Preço unitário do livro.	10.00
3	ValorBruto	Decimal(19,2)	A (sum)	Valor bruto da venda.	20.00
4	PercentagemDesconto	Decimal(19,2)	N	Valor percentual do desconto efetuado.	0.0
5	ValorDesconto	Decimal(19,2)	A (sum)	Valor do desconto efetuado sobre a venda.	0.00
6	ValorLíquido	Decimal(19,2)	A (sum)	Valor líquido da venda.	20.00
7	PercentagemLucro	Decimal(19,2)	N	Valor percentual do lucro da venda.	5.0
8	ValorLucro	Decimal(19,2)	A (sum)	Valor do lucro realizado com a venda.	1.00
9	EmStocknaEncomenda	Inteiro	A (avg)	Número de exemplares do livro em stock na altura da venda.	4
10	DiasAtéEnvio	Inteiro	A (avg)	Número de dias passados até ao envio da venda realizada – (<u>DtEnvio</u> – <u>DtEncomenda</u>)	0
11	DiasAtéRecepção	Inteiro	A (avg)	Número de dias passados até à recepção da venda pelo cliente – (<u>DtRecepção</u> – <u>DtEncomenda</u>)	0



Uma Tabela de Factos

Índices			
Nr	Identificação	Tipo	Descrição
1	(LivroId, Vendald, Lojald, Funcionáriold, DtEncomenda, DtVenda, DtEnvio, DtRecepção PaísId, ClientId, GrupoAutoresId, Línguald, Génerold, Editoriald, LivroId)	Primário	Único, ordenado fisicamente (<i>clustered</i>) de forma crescente.
2	Lojald	Secundário	Ordenado de forma crescente.
3	Funcionáriold	Secundário	Ordenado de forma crescente.
4	DtVenda	Secundário	Ordenado de forma crescente.
5	ClientId	Secundário	Ordenado de forma crescente.
6	LivroId	Secundário	Ordenado de forma crescente.
Perfis de Utilização			
Administradores gerais, gestores de loja e gestores comerciais.			
Observações			
Todos os valores considerados nos atributos medida são em Euros (€). Qualquer valor relativo a uma venda de uma loja da "L&LNet" que não esteja situada num país da zona Euro deverá ser convertido para Euro antes de ser introduzido na tabela de factos.			
Versão 1.00/2012, Belo, O.			



O Tipo de uma Tabela de Factos

- O tipo de uma tabela de factos releva-nos algumas características muito pertinentes de uma tabela de factos, nomeadamente:
 - a forma como é povoada;
 - o tipo de elementos de dados irá acolher;
 - a forma como os seus dados serão extraídos das fontes de informação;
 - a cadência do seu povoamento;
 - a periodicidade de refreshamento.
- Existem três formas de caracterizar o tipo de uma tabela factos, nomeadamente: **transacional**, **instantâneo** e **acumulativo**.



7

Dimensões e Perspetivas de Análise

Caracterização Base de uma Dimensão
Definição e Caracterização de Hierarquias
Uma Dimensão Universal

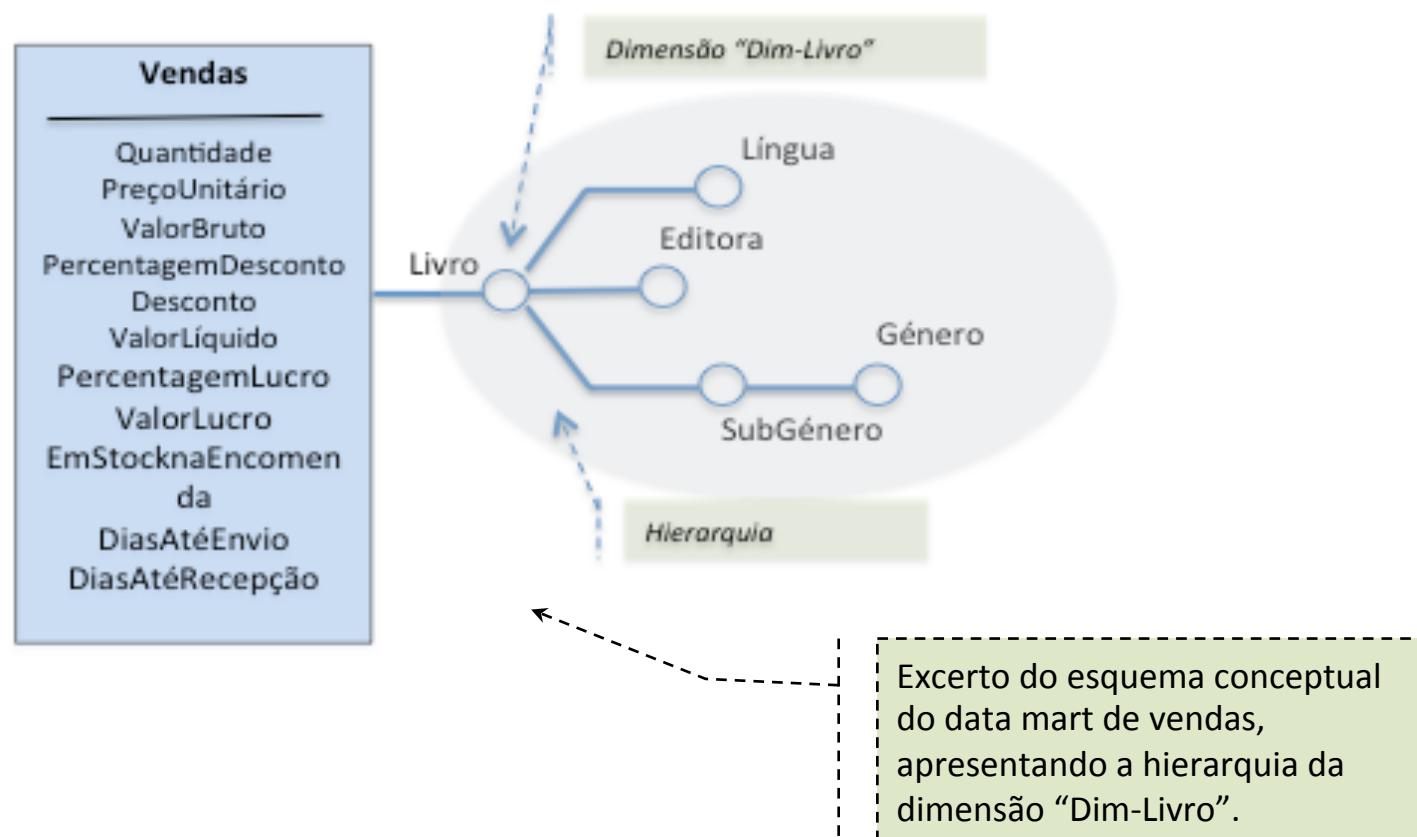


As Dimensões

- Uma dimensão suporta uma dada perspectiva de análise, categorizando um dado objeto de dados que pode ter uma ou mais referências em tabelas de factos.
- Uma dimensão disponibiliza um meio privilegiado de exploração das várias medidas de uma tabela de factos segundo as diversas perspectivas de análise que os seus atributos possibilitam.
- Os atributos de uma dimensão são definidos de acordo com as necessidades apresentadas pelos agentes de decisão.



O Modelo de uma Dimensão



Caracterização de uma Dimensão

Caracterização de dimensão							
Identificação		Dom-Livro					
Descrição		Caracterização base dos livros à venda na "L&L", com informação de negócio e apreciações críticas sobre os seus conteúdos					
Tipo		Com variação					
Dimensão		150KR					
Crescimento		2.5%/Ano					
Atributos							
Nº	Identificação	Descrição	Chave (Tipo)	Domínio (Tamanho)	V/H/P	Variação	Exemplos
1	LivroId	Código interno para a identificação do livro.	S	Inteiro	---	---	1
2	Titulo	Titulo do livro.	A	String(75)	---	---	The Data Warehouse Lifecycle Toolkit : Expert Methods for Designing, Developing, and Deploying Data Warehouses
3	Páginas	Número de páginas do livro.	N	Inteiro	---	---	568
4	Edição	Número da edição do livro.	N	String(10)	---	---	1ª edição
5	Capa	Descrição da capa do livro.	N	String(150)	---	---	Mole, de características genéricas
6	ImagenCapa	Imagen da capa do livro.	N	String(150)	---	---	\Livros\Capas\Imagens\1.jpg
7	DataPublicação	Data em que o livro foi publicado	N	Data	---	---	13/08/1998
8	ISBN-10	Indicação do código ISBN-10.	A	String(10)	---	---	0471255475
9	ISBN-13	Indicação do código ISBN-13.	A	String(15)	---	---	978-0471255475
10	Autores	Nomes dos autores do livro.	N	String(250)	---	---	Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W.
11	Editora	Editora do livro.	A	String(75)	---	---	Wiley
12	Língua	Língua na qual o livro foi escrito.	N	String(75)	---	---	Inglês
13	SubGénero	Subgênero literário no qual o livro se enquadra.	N	String(75)	---	---	Informática
14	Género	Gênero literário no qual o livro se enquadra.	N	String(75)	---	---	Técnico

- Caracterização geral.
- Atributos.
- Índices.
- Hierarquias.
- Perfis de utilização.



Caracterização de uma Dimensão

Caracterização de dimensão	
Identificação	Dim-Livro
Descrição	Caracterização base dos livros à venda na "L&L", com informação de negócio e apreciações críticas sobre os seus conteúdos
Tipo	Com variação
Dimensão	150KR
Crescimento	2.5%/Ano



Caracterização de uma Dimensão

Atributos							
Nº	Identificação	Descrição	Chave (Tipo)	Domínio (Tamanho)	V/H/P	Variação	Exemplos
1	LivroId	Código interno para a identificação do livro.	S	Inteiro	---	---	1
2	Titulo	Titulo do livro.	A	String(75)	---	---	The Data Warehouse Lifecycle Toolkit : Expert Methods for Designing, Developing, and Deploying Data Warehouses
3	Páginas	Número de páginas do livro.	N	Inteiro	---	---	568
4	Edição	Número da edição do livro.	N	String(10)	---	---	1ª edição
5	Capa	Descrição da capa do livro.	N	String(150)	---	---	Mole, de características genéricas
6	ImagenCapa	Imagem da capa do livro.	N	String(150)	---	---	\Livros\Capas\Imagens\1.jpg
7	DataPublicação	Data em que o livro foi publicado	N	Data	---	---	13/08/1998
8	ISBN-10	Indicação do código ISBN-10.	A	String(10)	---	---	0471255475
9	ISBN-13	Indicação do código ISBN-13.	A	String(15)	---	---	978-0471255475
10	Autores	Nomes dos autores do livro.	N	String(250)	---	---	Kimball, R., Reeves, L., Ross, M., & Thorntwaite, W.
11	Editora	Editora do livro.	A	String(75)	---	---	Wiley
12	Língua	Língua na qual o livro foi escrito.	N	String(75)	---	---	Inglês



Caracterização de uma Dimensão

Índices					
Nº	Identificação	Índice	Tipo		
1	LivroId	Primário	Único, ordenado fisicamente (<i>clustered</i>) de forma crescente.		
2	Título	Secundário	Ordenado de forma crescente.		
3	ISBN-10	Secundário	Único, ordenado de forma crescente.		
4	ISBN-13	Secundário	Único, ordenado de forma crescente.		
5	Editora	Secundário	Ordenado de forma crescente.		
Hierarquia (Ramos)					
Nº	Identificação	Esquema			
1	H1	LivroId -> Título -> Editora -> ALL			
2	H2	LivroId -> Título -> Língua -> ALL			
3	H3	LivroId -> Título -> Subgênero -> Gênero -> ALL			
Perfis de Utilização					
Administradores gerais, gestores de loja e gestores comerciais.					
Observações					
			Versão 1.00/2012, Belo, O.		
			<input type="checkbox"/>		



As Hierarquias

- A definição e a caracterização das hierarquias numa dimensão permite indicar os caminhos de agregação-desagregação que se podem seguir (GROUP BY paths) a partir de um dado nível de um dado atributo de uma dimensão. Uma hierarquia define uma sequência lógica de atributos.
- Uma hierarquia (Golfarelli e Rizzi, 2009) é uma árvore direcionada na qual os nodos são constituídos por atributos de dimensões e os ramos representam relacionamento do tipo muitos-para-um (N:1) entre pares de atributos dimensionais.
- Por exemplo, os ramos da hierarquia:
 - H1: LivroId -> Título -> Editora -> ALL
 - H2: LivroId -> Título -> Língua -> ALL
 - H3: LivroId -> Título -> Subgénero -> Género -> ALL



Tipos de Hierarquias

- **Simples ou estritas** (*strict*), na qual cada um dos nodos (ou níveis) apenas tem um nodo ascendente (“pai”) e um nodo descendente (“filho”), tendo contudo o mesmo critério de análise.
- **Não estritas** (*non-strict*), na qual cada nível na árvore da hierarquia tem apenas um possível nível pai, tendo o mesmo critério de análise; cada instância pode, contudo, ter mais do que uma instância no nível pai.
- **Múltipla e de caminhos alternativos** (*multiple and alternate path*), na qual cada nível da árvore da hierarquia pode ter mais do que uma instância no nível pai.
- **De caminhos paralelos** (*parallel path*), na qual cada nível da árvore poderá ter mais do que um nível pai e cada instância de um dado nível poderá pertencer a mais do que uma instância no nível pai.



Uma Dimensão Universal

Caracterização de dimensão																			
Nº	Identificação	Descrição	Chave (Tipo)	Domínio (Tamanho) [Range]	V/H/P	Variação	Exemplos												
1	DataId	Data do calendário	S	Data	---	---	30/12/2011												
2	DiaSemana	Numero do dia da semana	N	Inteiro	---	---	6												
3	NomeDiaSemana	Nome do dia da semana		String(15)	---	---	Sexta-feira												
4	Semana	Número da semana	A	Inteiro	---	---	52												
5	DiaMês	Número do dia do mês	N	Inteiro	---	---	30												
6	Mês	Número do mês	N	Inteiro	---	---	12												
7	NomeMês	Nome do mês		String(15)	---	---	Dezembro												
8	Trimestre	Número do trimestre	N	Inteiro	---	---	4												
9	Semestre	Número do semestre	N	Inteiro	---	---	2												
10	DiaAno	Número do dia do ano	N	Inteiro	---	---	364												
11	Ano	Número do ano	N	Inteiro	---	---	2011												
12	Feriado	Indicação se é ou não um dia feriado	N	String(1)	---	---	N												
13	EstaçãoAno	Nome da estação do ano	N	String(25)	---	---	Inverno												
14	EventosGrid	Eventos que ocorrem nesta data	N	Inteiro	---	---	1												
Índices																			
Nº	Identificação	Índice	Tipo																
1	DataId	Primário		Único, ordenado fisicamente (<i>clustered</i>) de forma crescente.															
2	Semana	Secundário		Ordenado de forma crescente.															
Hierarquia (Ramos)																			
Nº	Identificação	Esquema																	
1	H1	DataId -> Mês -> Trimestre -> Semestre -> Ano -> ALL																	
2	H2	DataId -> Feriado -> ALL																	
3	H3	DataId -> DiaSemana -> Semana -> Ano -> ALL																	
4	H4	DataId -> Estação -> Ano -> ALL																	
Perfis de Utilização																			
Administradores gerais, gestores de loja e gestores comerciais.																			
Observações																			
Nada a assinalar.																			
Versão 1.00/2012, Belo, O.																			

- Caracterização geral.
- Atributos.
- Índices.
- Hierarquias.
- Perfis de utilização.



Uma Dimensão Universal

Caracterização de dimensão							
Identificação		Dim-Calendário					
Descrição		Calendário do ano e seus atributos					
Tipo		Com diferentes papéis (<i>role-playing dimension</i>)					
Dimensão		3.7KR (Registos gerados durante o primeiro povoamento da dimensão)					
Crescimento		Não cresce. O povoamento desta dimensão é feito durante a fase de arranque do data warehouse para um período de 10 anos					
Atributos							
Nº	Identificação	Descrição	Chave (Tipo)	Domínio (Tamanho) [Range]	V/H/P	Variaçāo	Exemplos
1	DataId	Data do calendário	S	Data	---	---	30/12/2011
2	DiaSemana	Numero do dia da semana	N	Inteiro	---	---	6
3	NomeDiaSemana	Nome do dia da semana		String(15)	---	---	Sexta-feira
4	Semana	Número da semana	A	Inteiro	---	---	52
5	DiaMês	Número do dia do mês	N	Inteiro	---	---	30
6	Mês	Número do mês	N	Inteiro	---	---	12
7	NomeMês	Nome do mês		String(15)	---	---	Dezembro
8	Trimestre	Número do trimestre	N	Inteiro	---	---	4
9	Semestre	Número do semestre	N	Inteiro	---	---	2
10	DiaAno	Número do dia do ano	N	Inteiro	---	---	364
11	Ano	Número do ano	N	Inteiro	---	---	2011
12	Feriado	Indicação se é ou não um dia feriado	N	String(1)	---	---	N
13	EstaçãoAno	Nome da estação do ano	N	String(25)	---	---	Inverno
14	EventosGrid	Eventos que ocorrem nesta data	N	Inteiro	---	---	1



Uma Dimensão Universal

Índices					
Nr	Identificação	Índice	Tipo		
1	Datald	Primário	Único, ordenado fisicamente (<i>clustered</i>) de forma crescente.		
2	Semana	Secundário	Ordenado de forma crescente.		
Hierarquia (Ramos)					
Nr	Identificação	Esquema			
1	H1	Datald --> Mês -> Trimestre -> Semestre -> Ano -> ALL			
2	H2	Datald -> Feriado -> ALL			
3	H3	Datald -> DiaSemana -> Semana -> Ano -> ALL			
4	H4	Datald -> Estação -> Ano -> ALL			
Perfis de Utilização					
Administradores gerais, gestores de loja e gestores comerciais.					
Observações					
Nada a assinalar.					
			Versão 1.00/2012, Belo, O.		



Tipificação Geral

Dimensões com Variação – *Slowly Changing Dimensions*

Dimensões Conforme ou Partilhadas – *Conformed Dimensions*

Dimensões Degeneradas – *Degenerate Dimension*

Mini-dimensões

Dimensões Gigantescas – *Huge Dimensions*

Subdimensões – *Outriggers*

Dimensões com Diferentes Papéis – *Role-Playing Dimensions*

Dimensões de Controlo – *Junk Dimensions*

Dimensões de Origem Externa

8

Tipos de Dimensões e suas Variantes



Tipificação Geral

- As dimensões podem ser categorizadas a um nível superior de acordo com o tipo de estrutura que têm, com as operações de manipulação que sofrem ou com o seu número de atributos ou de registos que contêm, entre outras coisas. Os tipos que catalogámos foram os seguintes:
 - com variação (*slowly or rapidly changing dimension*);
 - conforme ou partilhada (*conformed dimension*);
 - degenerada (*degenerate dimension*);
 - gigantesca (*huge dimension*);
 - mini dimensão (*mini dimension*);
 - sub dimensão (*outrigger, outboard, or reference dimension*);
 - com diferentes papéis (*role-playing dimension*);
 - de controlo ou auxiliar (*junk dimension*).
 - de origem externa (*outsourced dimension*).



Os Tipos das Dimensões

- A categorização de uma dimensão está muito dependente do tipo de serviço que esta presta ao SDW durante as suas fases de povoamento, refletindo:
 - a forma como os dados vão evoluindo ao longo da vida do *data warehouse*;
 - a forma como é realizada a exploração de dados (sub dimensões, mini dimensões, dimensões com diferentes papéis, etc.);facilitando os processos de interrogação e as preferências de exploração de dados de um ou mais perfis de utilização.



Dimensões com Variação

- As tabelas de dimensão apresentam uma característica que as diferencia de todas as outras tabelas existentes num *data warehouse*:
 - os valores dos seus atributos podem variar ao longo do tempo.
- Isto contrasta claramente com uma das características base de um *data warehouse*: **a sua não volatilidade** (os dados são factuais e como tal não devem ser atualizados, apenas podem ser lidos).



Dimensões com Variação

- Numa dimensão, os atributos (e consequentemente os seus valores) foram escolhidos **para fazer a sua caracterização**, suportar as diferentes perspetivas de exploração de dados sobre a dimensão.
- As dimensões **não constituem qualquer elemento factual** relacionado com qualquer uma das áreas de decisão em que estão envolvidas, fornecendo elementos de dados que comprovem esta ou aquela situação
 - “simplesmente” atuam como **elementos de descrição, de filtragem ou de agregação. Mas...**



Dimensões com Variação

- (Kimball e Ross, 2002) categoriza as dimensões com variação (*slowly changing dimensions*), essencialmente em quatro tipos :
 - Tipo 1 – Reescrita simples dos valores afetados.
 - Tipo 2 – Criação de novos registo na tabela base.
 - Tipo 3 – Criação de novos atributos.
 - Tipo 4 – Criação de tabelas de histórico.

E alguns outros mais (0, 5,6 ou 7).



Dimensões com Variação T2 e T3

SCD T2													
SK	Clienteld	Nome	(...)	Idade	(...)	Localidadeld	Zonalld	(...)	Ranking	(...)	Datalnicial	DataFinal	
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	
23	23	Francisco Belarmino	(...)	25	(...)	1	1	(...)	2	(...)	2010/01/01	2011/02/03	
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	
45	SCD T3		Clienteld	Nome	(...)	Localidadeld	LocalidadeldA	DataActL	Zonalld	ZonalldA	DataActZ	(...)	
Um f	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	
23		João Francisco Belarmino	(...)	1		2		01/01/2010	1	NULL	NULL	(...)	
24		Carlos Sergio Ramirez	(...)	1		NULL		NULL	1	NULL	NULL	(...)	
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	

Um fragmento de uma dimensão com variação do tipo 3



Dimensões com Variação T4

SCD T4

Nome		(...)	Idade	(...)	LocalidadId	Zonald	(...)	Ranking	(...)
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)
23	João Franscisco Belarmino	(...)	25	(...)	1	1	(...)	2	(...)
24	Carlos Sergio Ramirez	(...)	32	(...)	1	1	(...)	57	(...)
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)

Clientid	HstNr	DataRef	(...)	Idade	(...)	LocalidadId	Zonald	(...)	Ranking	(...)
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)
10	44	01/12/2009	(...)	65	(...)	2	1	(...)	8	(...)
23	45	01/01/2010	(...)	25	(...)	1	1	(...)	6	(...)
44	46	05/08/2011	(...)	18	(...)	1	1	(...)	9	(...)
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)

Um fragmento de uma dimensão com variação do tipo 4

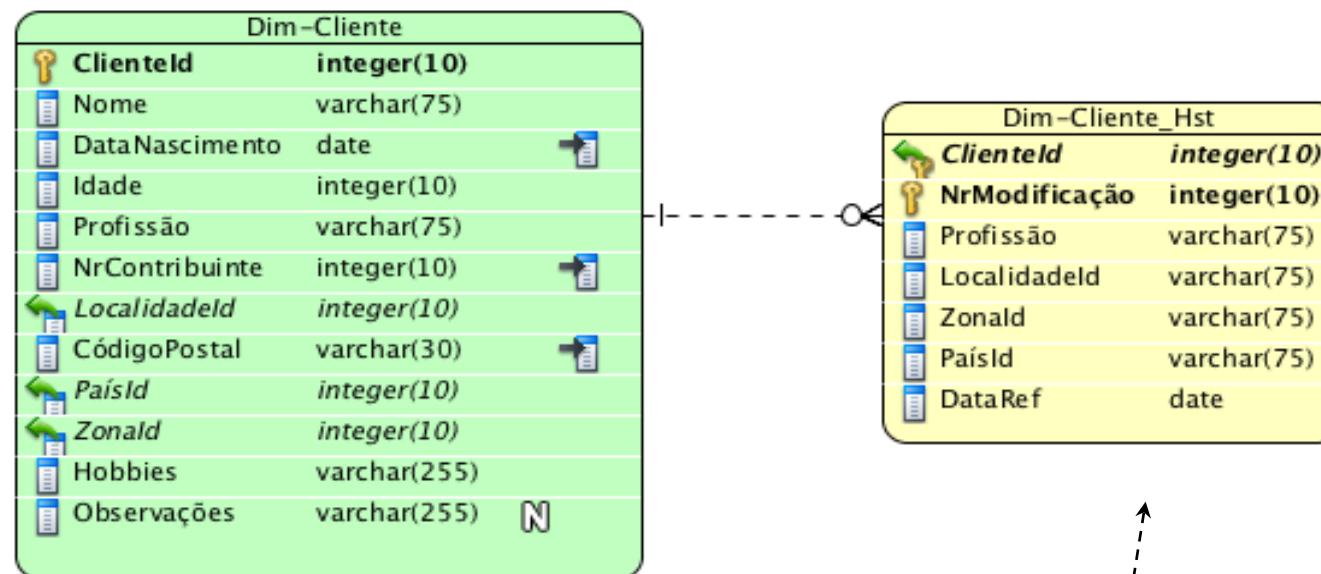


Uma Outra Tipificação

- Ao caracterizarmos uma dimensão com variação devemos indicar sempre, **para cada um dos seus atributos**, a forma como queremos que o futuro sistema de povoamento atue sobre cada um deles.
- Para isso devemos indicar, atributo a atributo, os seguintes aspectos:
 - Variação ('S'/'N').
 - História ('S'/'N').
 - Periodicidade ('?'/‘R’/‘D’/‘W’/‘M’/‘T’/‘S’/‘A’/‘I’).



Uma Dimensão com Variação T4



O Esquema da tabela de histórico deve ser definido com base os requisitos de variação identificados no início do projeto.

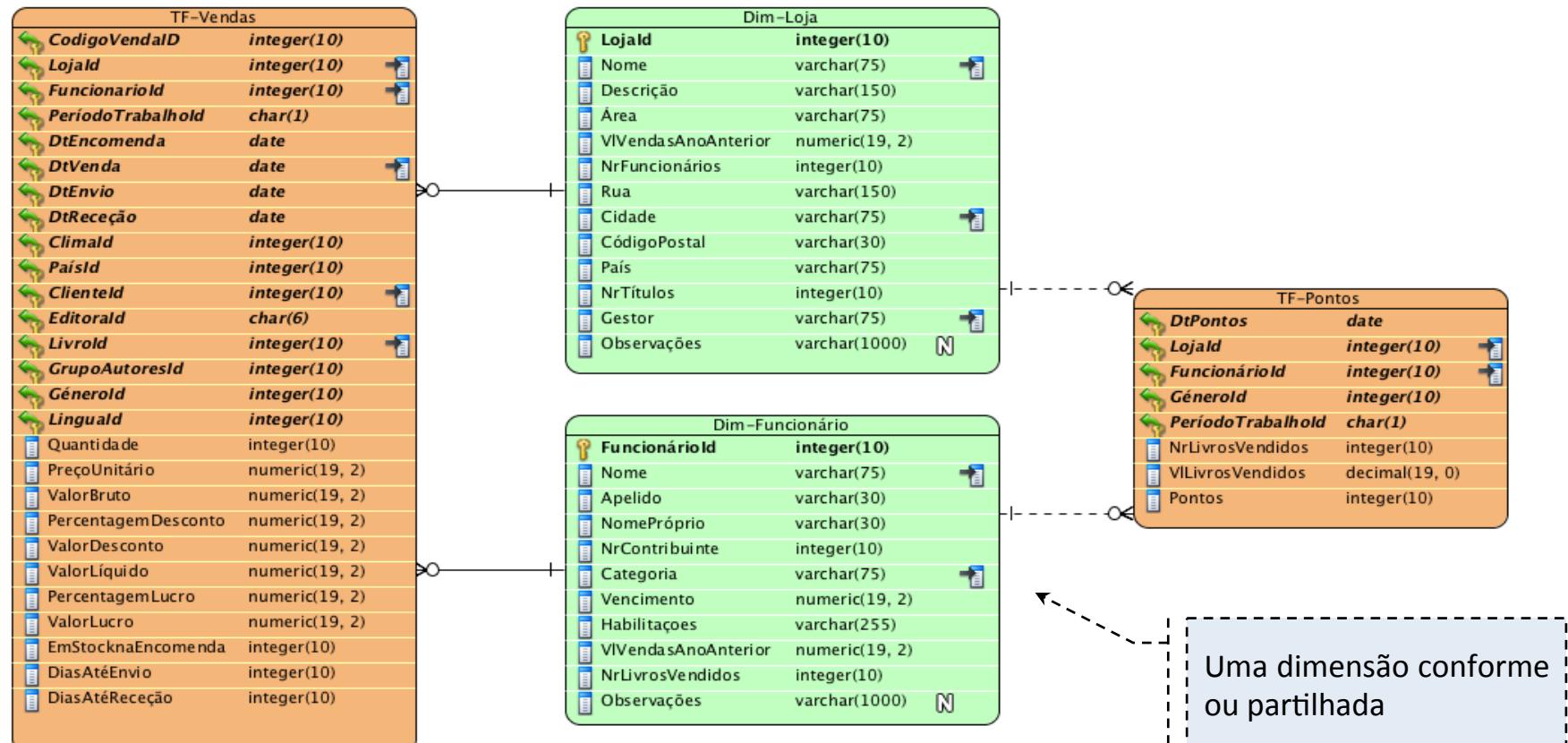


Dimensões Conforme

- Quando lidamos com mais do que uma área de decisão – vários esquemas para suporte a problemas diferentes - é possível que aconteça que **mais do que uma tabela de factos incorpore uma mesma dimensão na sua estrutura.**
- A não ser que, por motivos de desempenho ou de estruturação de dados se opte por definir duas tabelas separadas para a mesma dimensão – uma para cada tabela de factos -, a tabela que acolherá a informação relativa à dimensão em causa **será partilhada pelas duas tabelas de facto.**



Dimensões Conforme



Dimensões Degeneradas

- Uma dimensão degenerada (*degenerate dimension*) é uma dimensão que não possui qualquer atributo que a caracterize para além daquele que integra a chave da tabela de factos.
- Na prática, isto significa que é uma dimensão que não tem uma tabela de suporte autónoma para a sua operacionalização, tal como sucede com a generalidade dos tipos de tabelas de dimensão.
- A sua existência num esquema dimensional é revelada apenas por um (ou mais) atributo(s) integrado(s) na estrutura de uma tabela de factos.



Dimensões Degeneradas

TF-Vendas	
LohalID	integer(10)
VendaID	integer(10)
FuncionárioId	integer(10)
DtEncomenda	date
DtVenda	date
DtEnvio	date
DtReceção	date
ClienteId	integer(10)
AutorId	integer(10)
LivroId	integer(10)
Quantidade	integer(10)
PreçoUnitário	integer(10)
ValorBruto	integer(10)
Desconto	integer(10)
ValorLíquido	integer(10)
PercentagemLucro	integer(10)
ValorLucro	integer(10)
EmStocknaEncomenda	integer(10)
DiasAtéEnvio	integer(10)
DiasAtéReceção	integer(10)

“VendaID” (e “PontoDeVendaID”) é um atributo relativo a uma dimensão degenerada, uma vez que não tem uma tabela de dimensão associada.

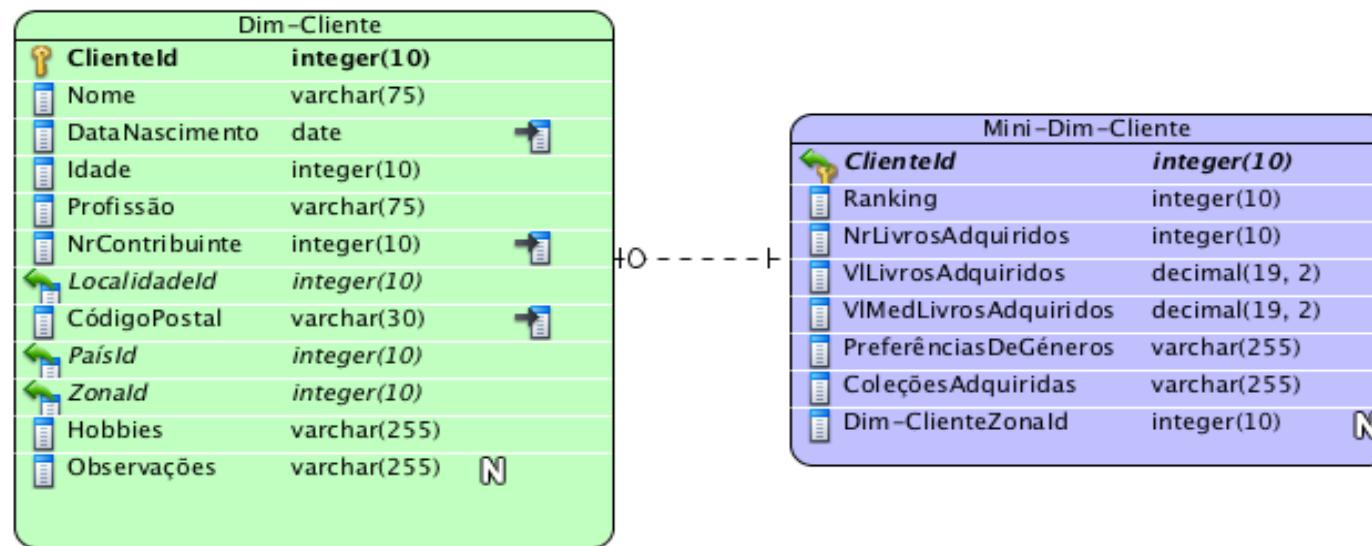


Mini-Dimensões

- Por vezes, na caracterização de uma dimensão integramos um número de atributos tal, que a sua estrutura atinge uma dimensão bastante significativa.
- Se fizermos um estudo sobre a exploração futura dos dados deste tipo de dimensões, verificamos que
 - uma grande maioria dos seus atributos raramente intervêm nas *queries* lançadas pelos agentes de decisão – o efeito de “lastro”.
 - existe um dado conjunto de atributos cujos valores variam com alguma frequência (lenta ou rápida) ao longo do tempo.



Mini-Dimensões



Exemplo de uma mini-dimensão e a sua correspondente dimensão base. A mini-dimensão poderá estar ligada diretamente a uma tabela de factos.



Dimensões Gigantescas

- A categorização de uma dimensão como sendo gigantesca (*huge ou large dimension*) é muito subjetiva, dependendo muito do contexto e das características do problema que temos em mãos.
- De acordo com o nosso conhecimento e experiência no domínio é “fácil” reconhecer se uma dimensão é ou não gigantesca - isso tem a ver, na generalidade dos casos, apenas com a nossa capacidade de a manipular com ou sem dificuldades.

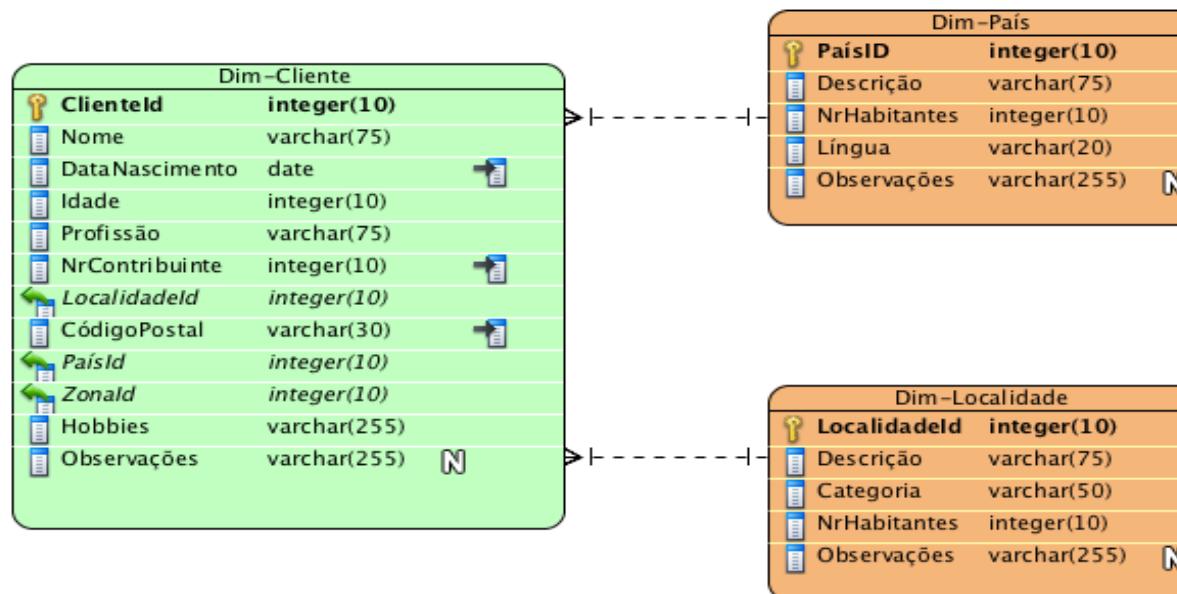


Sub-Dimensões (*Outriggers*)

- As sub-dimensões, também designadas por dimensões secundárias, de referência ou *outiggers*, são tabelas definidas especificamente para apoiarem a caracterização de uma dimensão designada como principal, de primeira linha, que num esquema dimensional está ligada diretamente a pelo menos uma tabela de factos.
- Uma sub-dimensão não desempenha um papel de primeiro plano nos processos de exploração de dados realizados sobre o esquema dimensional em que está localizada.
- Em geral, são tabelas criadas para reduzirem o nível de redundância existente na dimensão à qual estão ligadas, muitas das vezes provocado pela existência de simples dependências funcionais transitivas entre alguns dos atributos da dimensão.



Sub-Dimensões (*Outriggers*)

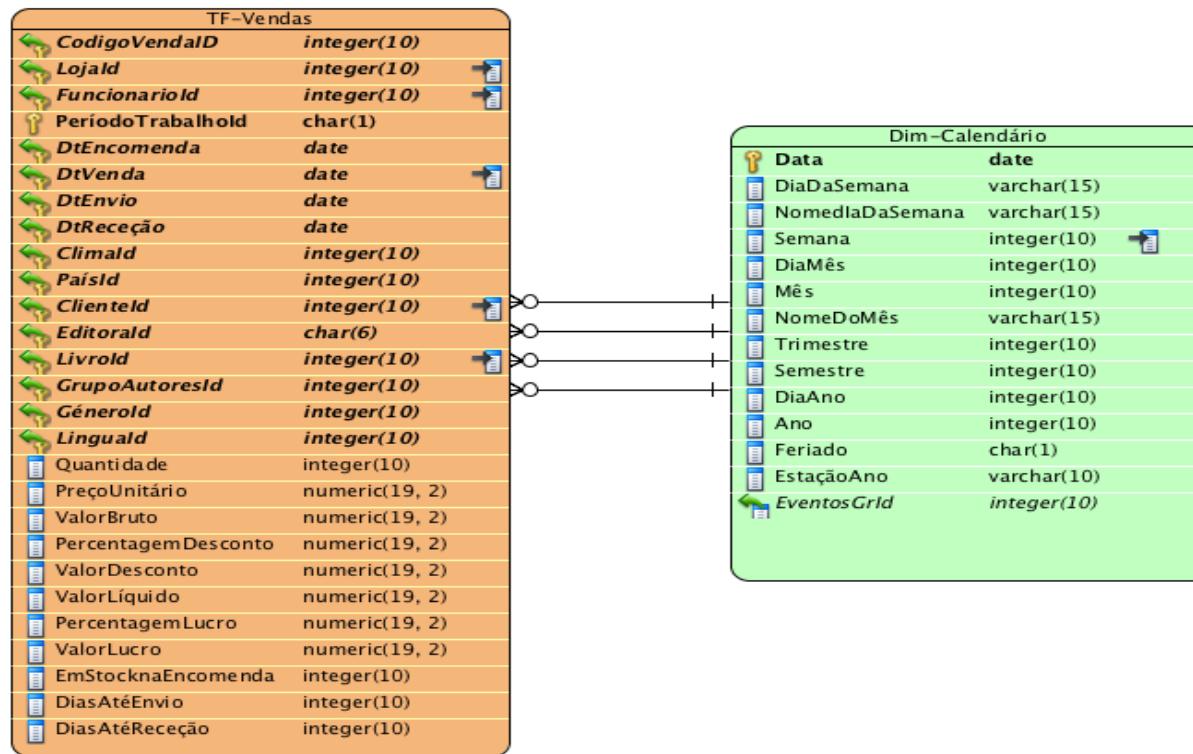


Dimensões com Diferentes Papéis

- Uma dimensão com diferentes papéis (*role-playing dimension*) é uma tabela de dimensão regular, mas que está referenciada mais do que uma vez numa mesma tabela de factos.
- Cada uma dessas referências, cada um desses atributos, dá sentido a um contexto de aplicação diferente, fazendo com que o atributo em questão tenha um significado distinto dos seus “semelhantes”, apesar da sua natureza ser a mesma.



Dimensões com Diferentes Papéis



Dimensões de Controlo

- Por vezes é necessário atuar no sentido de reduzir o tamanho de uma tabela de factos, porque:
 - pode estar a provocar algum **tipo de estrangulamento** no desempenho do sistema na satisfação das queries,
 - queremos **simplificar os processos de monitorização** e controlo desenvolvidos sobre as estruturas de dados implementadas
 - Precisamos de **assegurar contextos transacionais** associados com um dada tabela de factos;
 - se quer obter um pouco mais de **comodidade** nos processos de manipulação de dados mantidos sobre o esquema dimensional em causa.

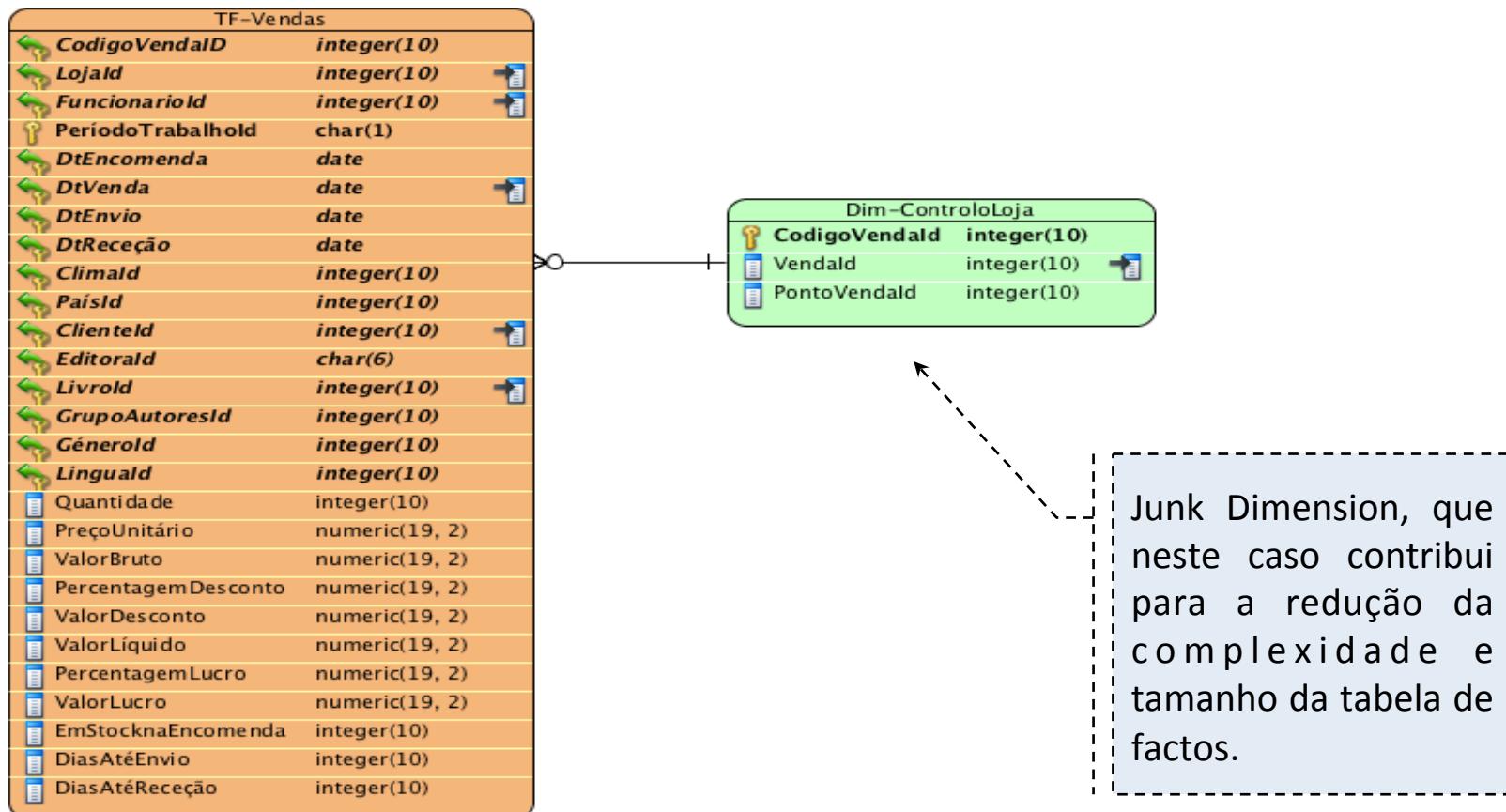


Dimensões de Controlo

- Uma dimensão de controlo (*junk dimension*) é pois uma dimensão abstracta especialmente concebida para acolher dados que não têm diretamente a ver com as atividades de análise subjacentes a um determinado esquema conceptual, mas cujos valores são muito importantes para os processos de tomada de decisão para que possam ser, simplesmente, desprezados.



Dimensões de Controlo

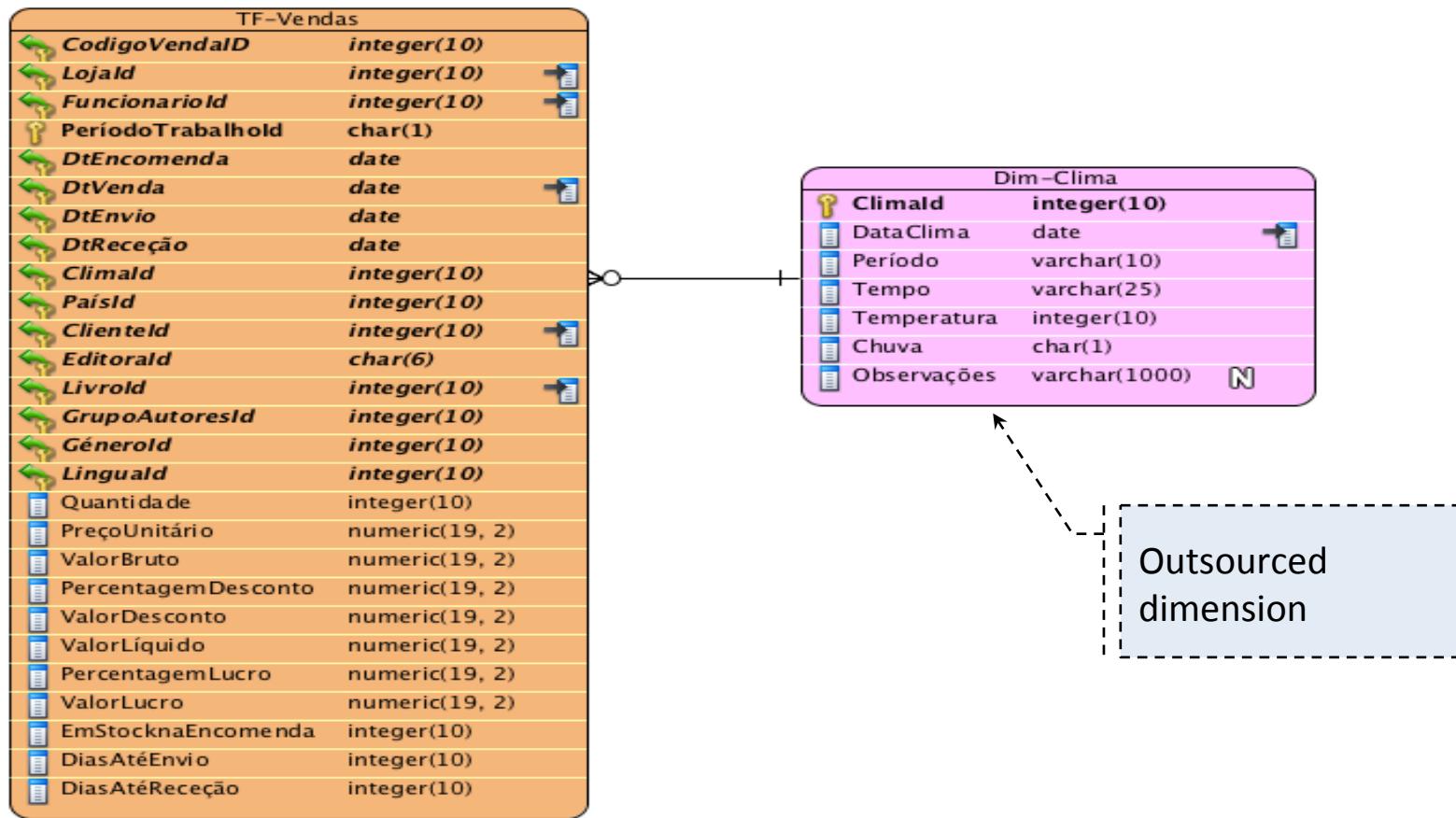


Dimensão de Origem Externa

- As dimensões de origem externa (*outsourced dimensions*) são assim designadas por materializar uma vertente de análise de negócio específica, que é suportada exclusivamente por um conjunto de atributos cujos valores são povoados a partir de fontes de informação externas à empresa.
- A forma como estas dimensões são geridas segue, em todos os aspetos, aquilo que é usual realizar-se para qualquer outra dimensão de análise com alimentação interna.



Dimensão de Origem Externa



Relacionamentos ‘muitos-para-muito’
Chaves de Substituição
Desnormalização de esquemas
Tabelas *Factless*

9 Casos Particulares



Relacionamentos N:M

- É vulgar (mas não desejável) aparecerem por vezes alguns casos de relacionamentos do tipo muitos-para-muitos (N:M) entre uma tabela de factos e uma tabela de dimensão.
 - É uma situação que ocorre quando vários registo de uma mesma dimensão estão relacionados ao mesmo tempo com um ou mais registo da própria tabela de factos.
- Na realidade, nada de novo em termos de relacionamento entre tabelas - num sistema operacional a solução seria óvia.

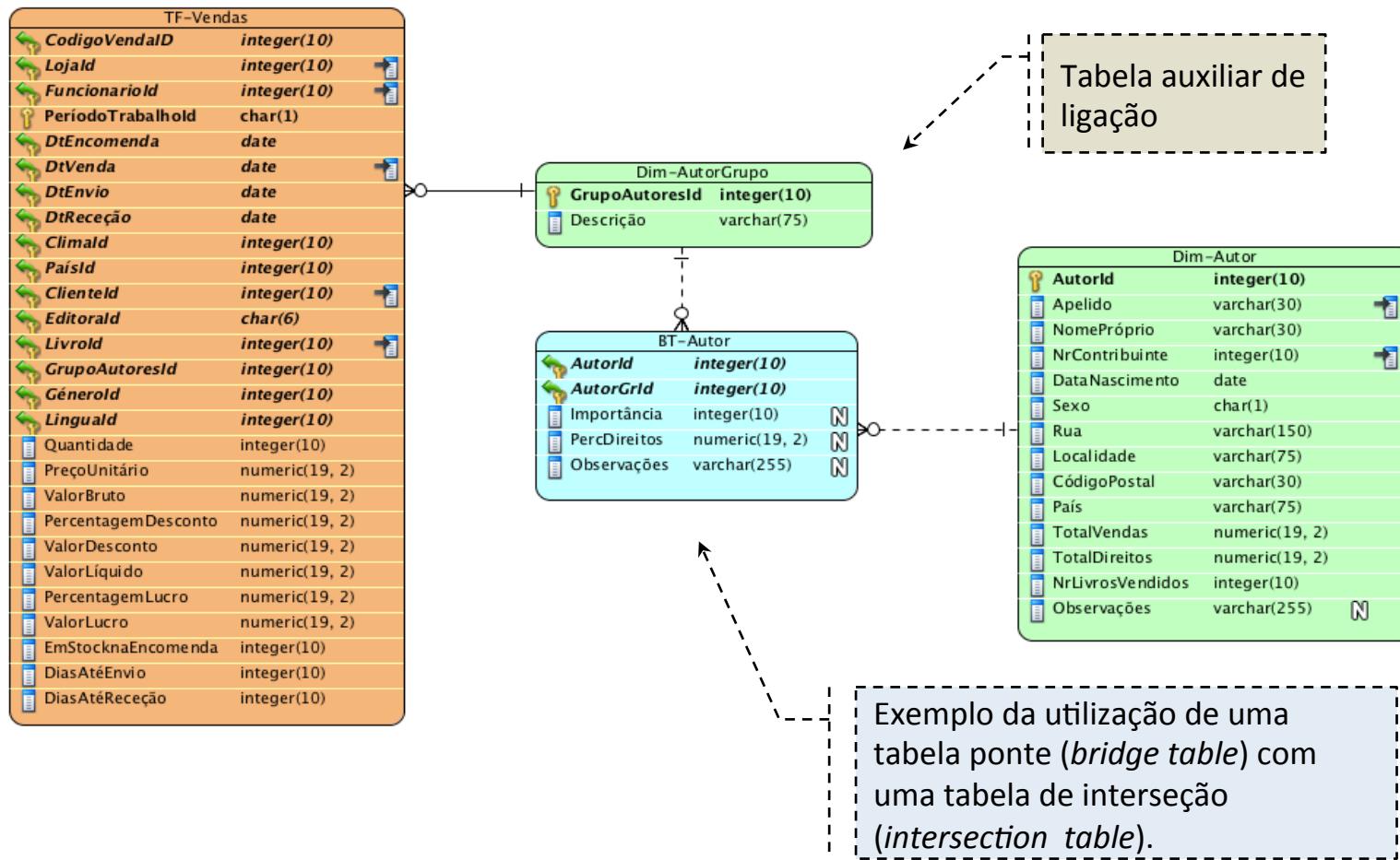


As Tabelas Ponte

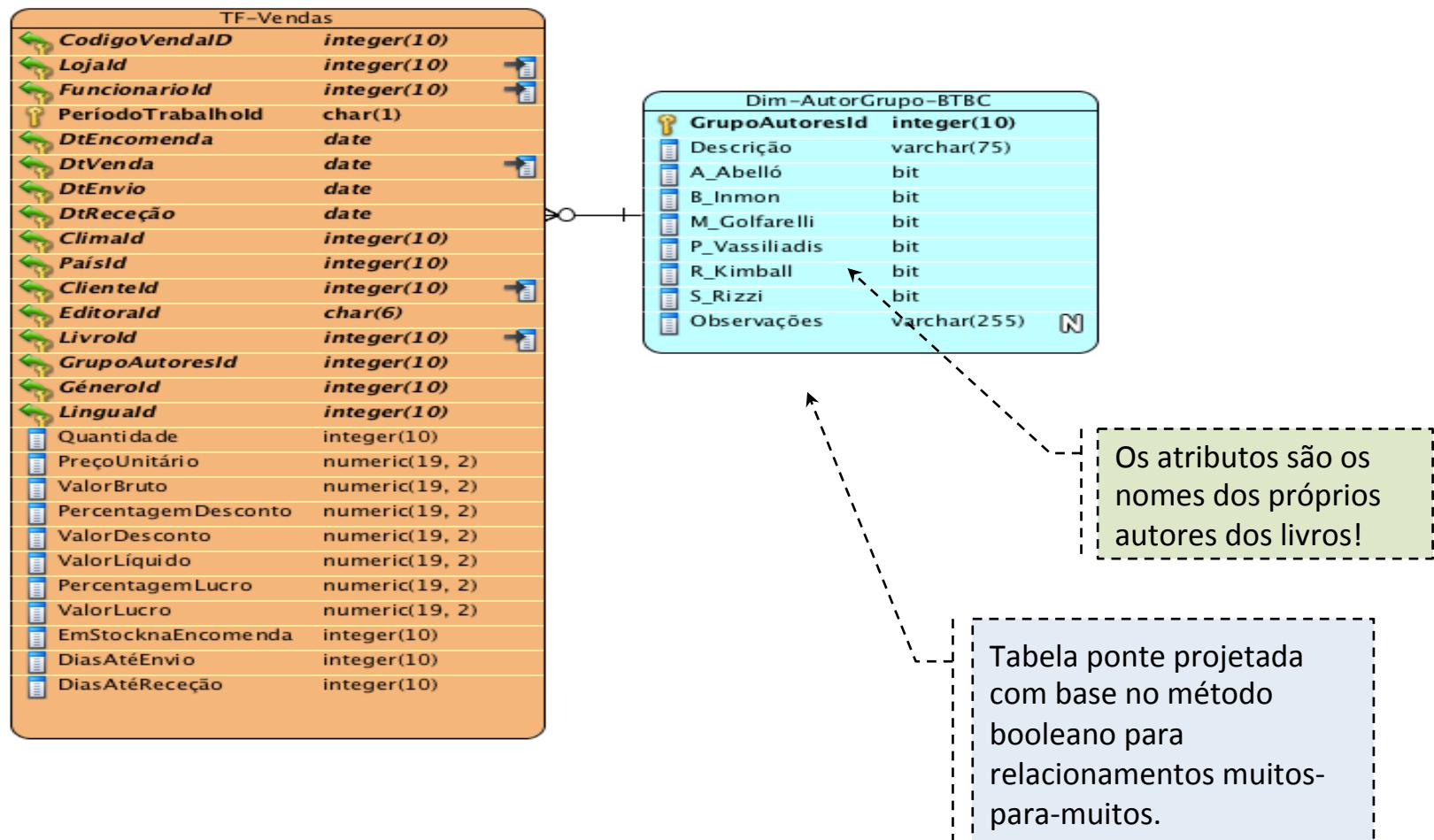
- Podemos estabelecer os relacionamento N:M de diferentes maneiras. Os métodos mais usuais utilizam como suporte (Kimball, 1998a) (Hamel, 2007) :
 - Uma tabela de interseção.
 - Um conjunto de atributos do tipo booleano.
 - Um conjunto de atributos para acolhimento de várias instâncias (múltiplas colunas).



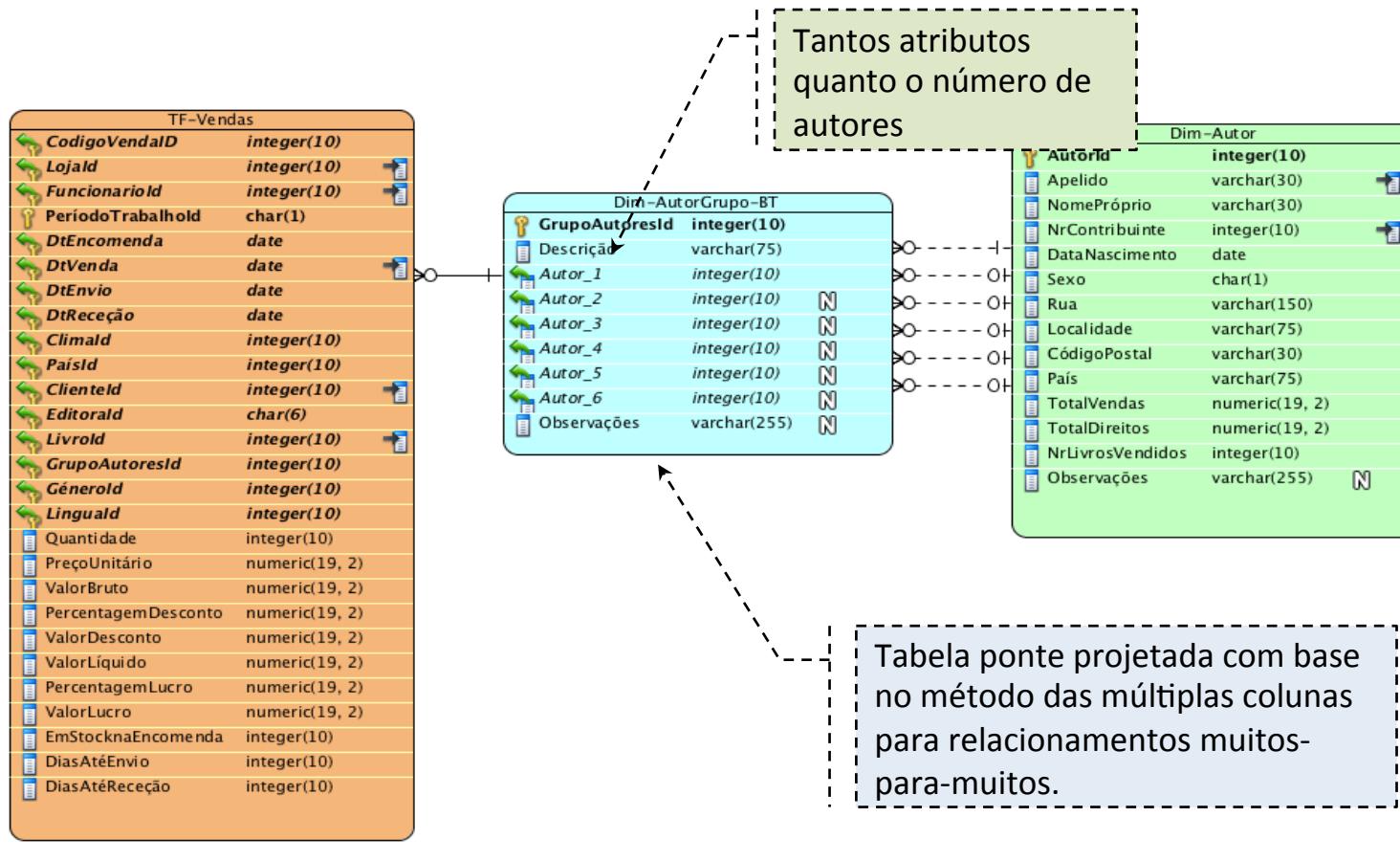
As Tabelas-Ponte – 1º Caso



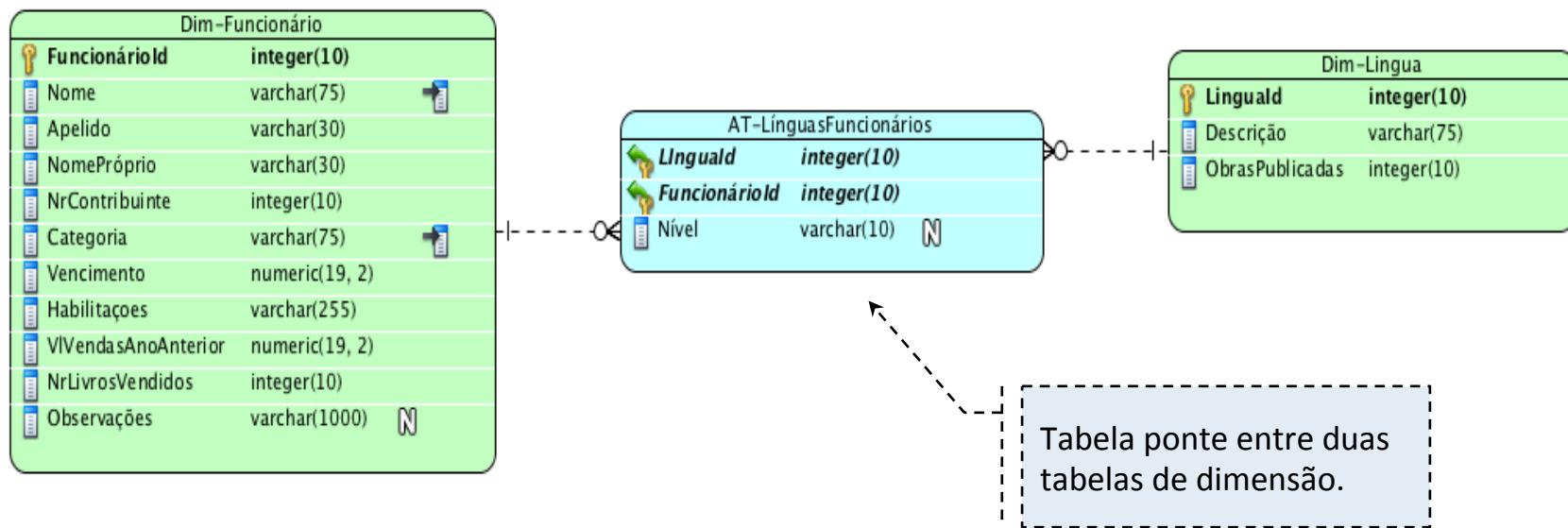
Tabelas-Ponte – 2º Caso



Tabelas-Ponte – 3º Caso



Tabelas de Associação



Tabelas de Factos *Factless*

- Poderão existir tabelas de factos sem qualquer métrica definida: as *factless* (“sem factos”) ou por tabela de junção (*junction table*). São por vezes utilizada em situações nas quais se pretende:
 - realizar intensivamente operações de contagem sobre um ou mais atributos de dimensão contidos na tabela de factos;
 - suportar a definição de relacionamentos muitos-para-muitos;
 - manter uma log de eventos.



Chaves de Substituição

- Nem sempre **as chaves operacionais** apresentadas pelas diversas fontes de informação podem ser adotadas nos *data warehouses*.
- Os **motivos** podem ser os mais diversos, mas na generalidade relacionam-se com:
 - a redução do espaço de armazenamento em disco;
 - a diminuição do tempo de satisfação das *queries*;
 - a heterogeneidade das fontes de informação.



Chaves de Substituição

- As **chaves de substituição (surrogate keys)** (Becker, 2006) são atributos criados especialmente com a finalidade de simplificar a forma de inter-relacionamento das tabelas de dimensão com as tabelas de factos, apresentando características como:
 - são constituídas por “pequenos” inteiros,
 - não relevam ou transmitem qualquer semântica acerca do negócio ou dos próprios processos
 - garantem a combinação adequada dos caminhos de junção em todos os processos em que atuam
 - têm um “custo” claramente inferior às suas congêneres naturais – se esquecermos o sistemas de ETL.



Operacionais vs Substituição

Tabela de Factos "TF-Vendas" Utilização de chaves de substituição						
Chave operacional	Espaço em disco (KBytes)	Cardinalidade diária	Espaço em disco (KBytes)	Chave de substituição	Espaço em disco (KBytes)	Percentagem da redução do espaço em disco
DtEncomenda	Date - 3 bytes	2.5KR/dia	7.5Kbytes	Smallint - 2 bytes	5Kbytes	33%
Editorald	Char(6) - 6 bytes	2.5KR/dia	15Kbytes	Tinyint - 1 byte	2.5Kbytes	83%
Clienteld	Integer(10) - 4 bytes	2.5KR/dia	10Kbytes	Int - 4 byte	10Kbytes	0%

Observações:
Para a elaboração deste estudo assumiu-se que todas as lojas da "L&LNet" registavam nos seus sistemas cerca de 2500 registo (2.5KR) por dia.

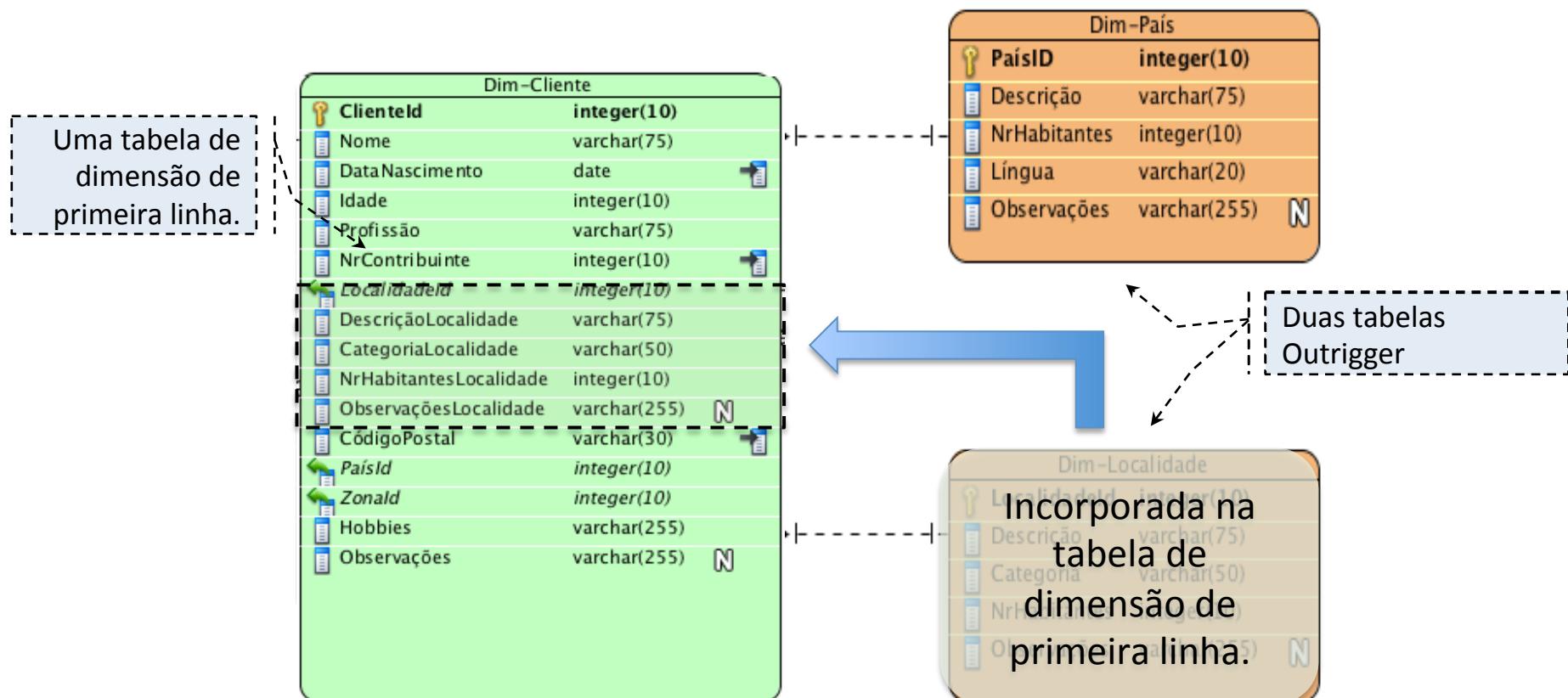


Desnormalização de Esquemas

- É frequente ouvir-se que o esquema dimensional de um *data warehouse* “é”, por natureza, desnormalizado. Mas isso não é verdade.
- Se procuramos conceber um esquema dimensional que proporcione **altos níveis de desempenho** na satisfação de *queries* é muito natural que se desnormalize partes significativas do seu esquema.
- Nos *data warehouses* a desnormalização de dados não é “**perigosa**”, uma vez que são repositórios de dados apenas de leitura.



Desnormalização de Esquemas



Que implicações terá este tipo de ação?



A Qualidade de um Esquema

- A **perícia e o conhecimento** do arquiteto de um *data warehouse* é muito importante na preparação e desenvolvimento de qualquer esquema dimensional.
- A qualidade do esquema dimensional, bem como a sua efetividade em termos de representação de dados, depende muito:
 - da forma como a área de suporte à decisão foi abordada,
 - dos requisitos dos seus agentes de decisão.



Estrelas e Flocos de Neve
Constelações e Galáxias

10 Configurações de Esquemas Dimensionais

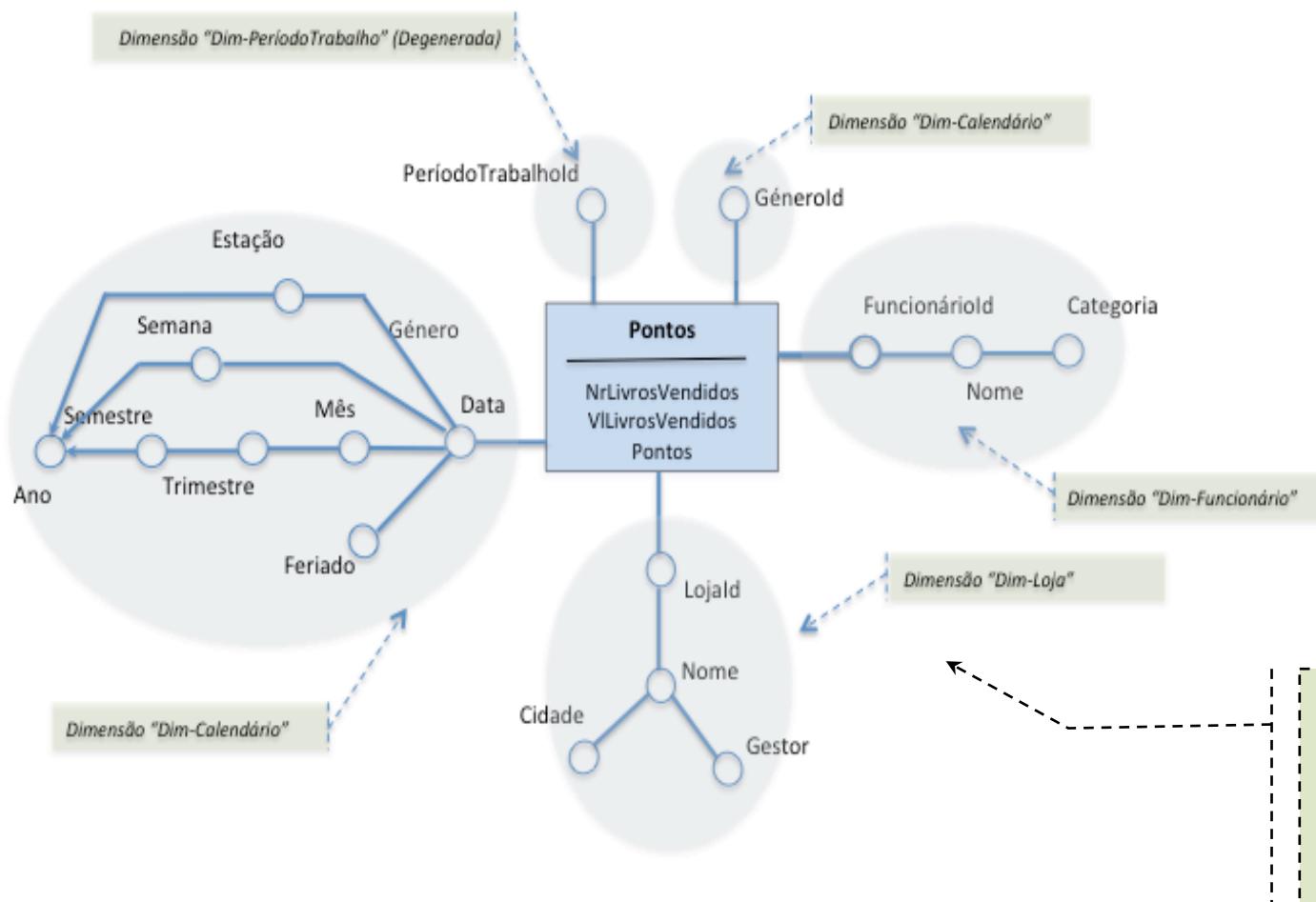


Estrelas e Flocos-de-Neve

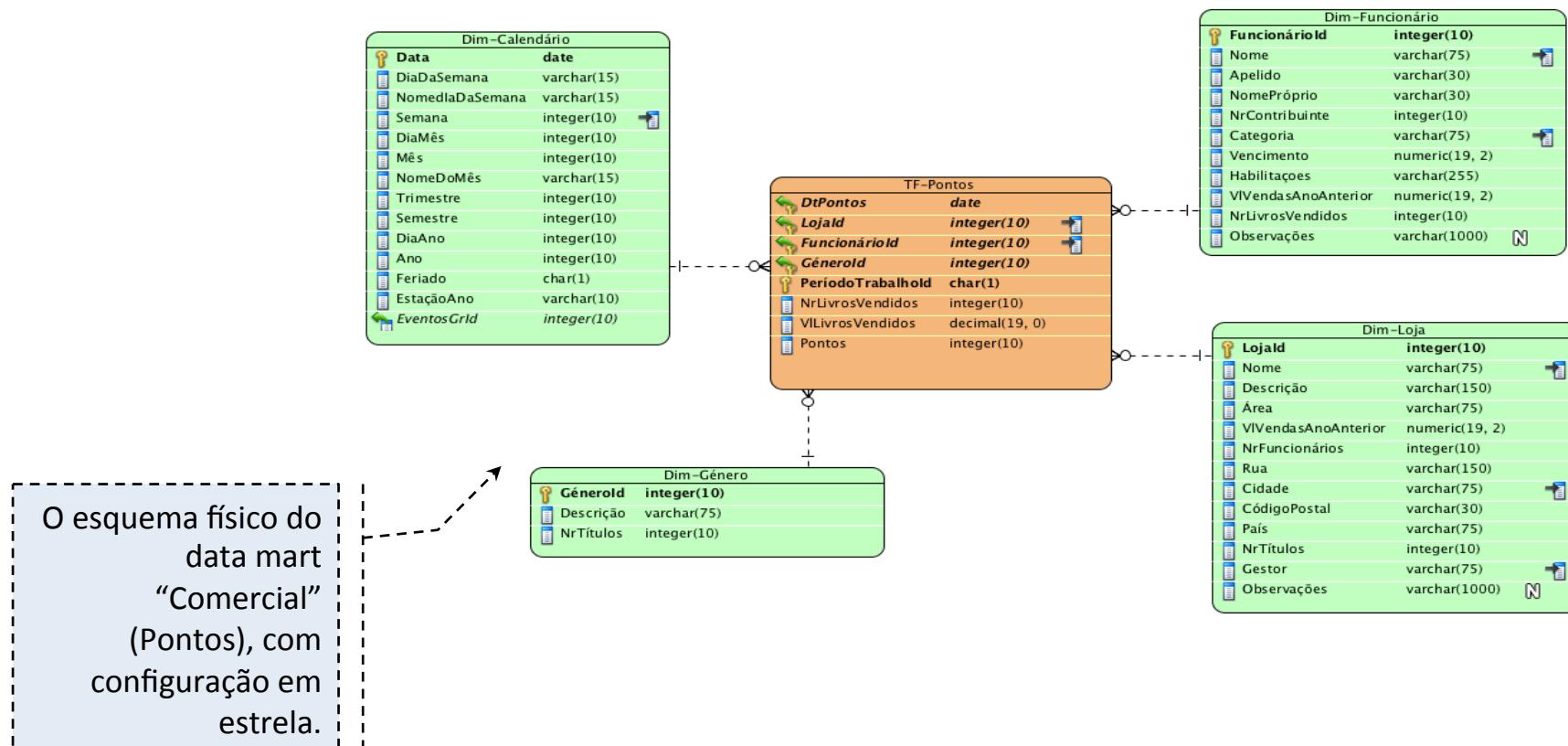
- Na terminologia de *data warehousing* encontramos frequentemente referências **designações um pouco particulares**, como estrelas, flocos de neve, constelações e algumas outras mais.
- A grande maioria delas é utilizada para **caracterizar a forma como as tabelas de factos e as dimensões de um esquema dimensional estão organizadas**.
- A configuração mais elementar (e a mais vulgar) é a configuração de um esquema dimensional em **estrela (*star schema*)**.



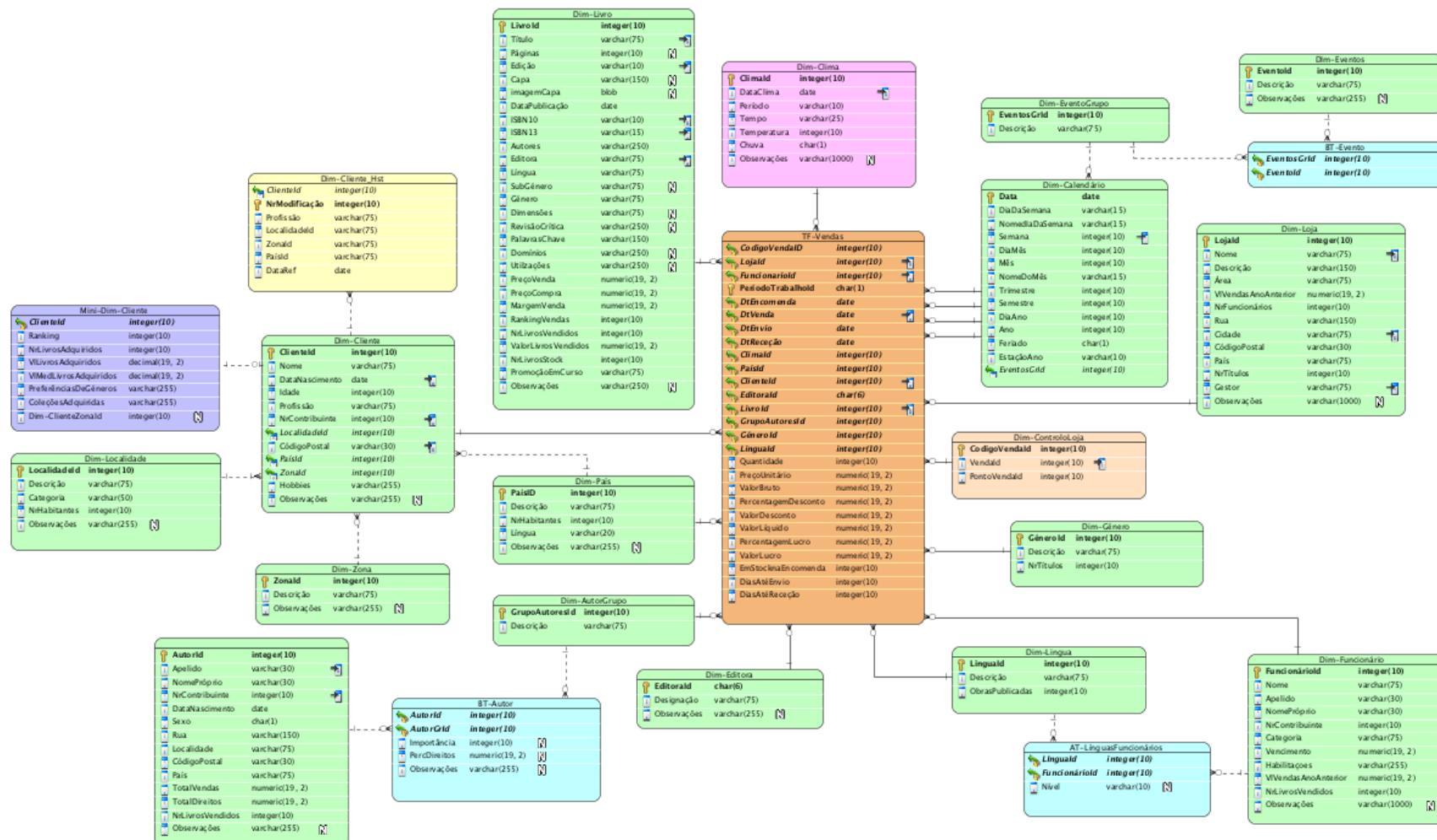
A Estrela dos Pontos



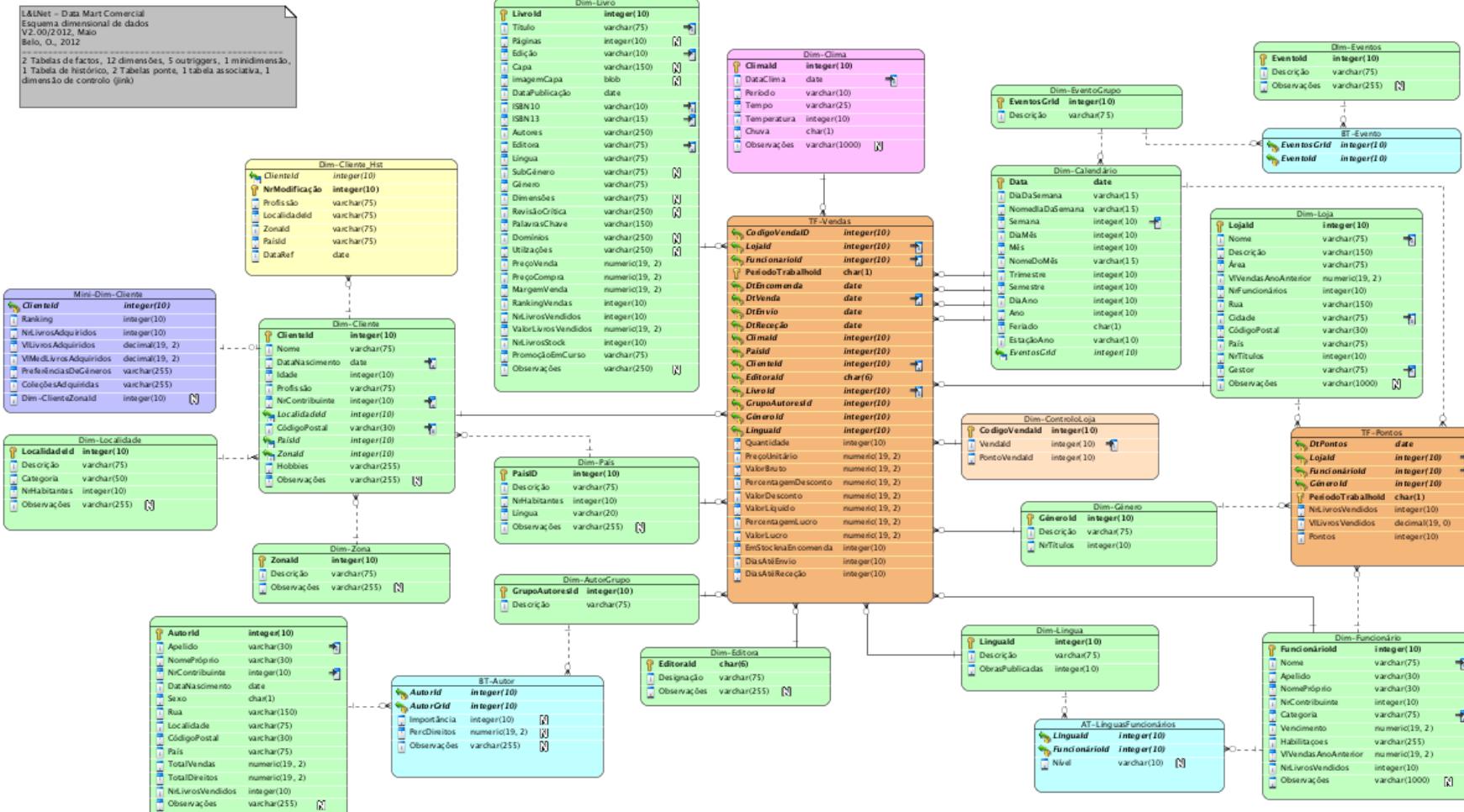
A Estrela dos Pontos



O Floco-de-Neve das Vendas



Constelações e Galáxias





11

Algumas Notas Finais

Comentários gerais sobre modelação dimensional

A Modelação Dimensional

- A atividade de **modelação dimensional** é **naturalmente complexa**, envolvendo aspectos que poderão ir desde simples questões para suporte ao **negócio** até à forma como o esquema final poderá ser abordado pelas **queries**.
- Projeto atrás de projeto os arquitetos de sistemas de *data warehousing* desenvolveram soluções muito expeditas, refletindo as mais variadíssimas necessidades dos **agentes de decisão** e, consequentemente, dos dados com que estes pretendem trabalhar.



Alguns Mitos

- Kimball e Ross (2002) enunciaram alguns “mitos” dos sistemas de *data warehousing*, nomeadamente, que os modelos dimensionais e os *data marts*:
 - são apenas desenvolvidos para dados sumariados (agregados);
 - são de âmbito departamental e não empresarial;
 - não são escaláveis;
 - apresentam um nível de integração tal que provocam estrangulamentos nas soluções adotadas.
 - (...)
- Com algum conhecimento e experiência prática rapidamente todos estes “mitos” são postos em causa.



Algumas Boas Práticas

- Margy Ross (2009) apresentou algumas recomendações (“**not-to-be-broken rules**”) que poderíamos designar como 10 regras “essenciais” para a modelação dimensional:
 1. Carregar as estruturas dimensionais com **dados atómicos** detalhados.
 2. Os modelos dimensionais devem ser **estruturados de acordo com os processos de negócio** e de tomada de decisão.
 3. Garantir que qualquer tabela de factos tenha associada **uma tabela de dimensão tempo – um calendário**.
 4. Garantir que todos os factos contidos numa única tabela de factos **respeitam o grão** definido, no mesmo nível de detalhe.
 5. **Eliminar relacionamentos muitos-para-muitos** em tabelas de factos.



Algumas Boas Práticas

6. Eliminar relacionamentos muitos-para-um nas tabelas de dimensão.
7. Armazenar nas tabelas de dimensão **etiquetas (labels)** para utilização em relatórios ou como valores de filtragem de domínios.
8. Ter a certeza que as tabelas de dimensão utilizam uma **chave de substituição (surrogate key)**.
9. Promover a criação de dimensões partilhadas (*conformed dimensions*) de forma a permitir a integração de dados em todos os domínios empresariais.
10. Equilibrar continuamente os requisitos e as realidades do negócio, de forma a providenciar uma solução que seja aceite pelos utilizadores e que possa suportar os seus processos de tomada de decisão.



A Nota Final

- A modelação dimensional de dados é uma atividade que se rege pelos **requisitos de processos de tomada de decisão** e não por requisitos de processos de suporte operacional.
 - **Esquecer isso** é “meio caminho andado” para o insucesso de um *data warehouse*.



Notas de Leitura em *Business Intelligence*

03 >> Modelação Dimensional de Dados

Orlando Belo

Departamento de Informática, Escola de Engenharia, Universidade do Minho

PORTUGAL

<fim>

