

SISTEMAS DE APOIO À INTELIGÊNCIA DE NEGÓCIOS

Asterio K. Tanaka

<http://www.uniriotec.br/~tanaka/SAIN>
tanaka@uniriotec.br



Modelagem Dimensional – Conceitos Avançados

Material baseado em originais de Maria Luiza Campos (<http://dataware.nce.ufrj.br/>)
Complementado com referências atuais de Ralph Kimball (<http://www.kimballgroup.com/>)
Agosto de 2007

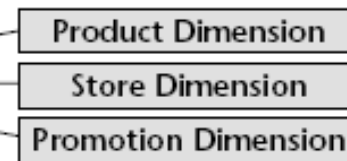
- Dimensões clássicas em negócio de varejo
 - When (Tempo, Data, Hora do Dia); What (Produto); Where (Loja); Why (Promoção)
- Tabelas de Fato sem Fatos ou Métricas
 - Cobertura (Promoção) e Evento
- Dimensões Degeneradas (dimensões sem tabelas)
- Extensibilidade do esquema estrela
- Modelo dimensional normalizado: Esquema Snow Flake
- Esquemas com muitas dimensões: Esquema Centípedo
- Campos Chaves de Tabelas de Dimensões
- Dinâmica das Dimensões: Slowly Changing Dimension
- Dimensões com Papéis (Role Playing dimensions)
- Outros Tipos Especiais de Dimensão
 - Lixo (Junk Dimension); Dimensões muito grandes: Minidimensões; Dimensões com “outrigger”; Dimensões Multivaloradas (Bridge table)
- Tópicos Especiais sobre Fatos
 - Fatos conformados, Bus Matrix de Implementação, Tipos Clássicos de Fatos
- Agregados

Dimensão tempo (Data)

- A dimensão tempo é muito poderosa e importante em todo DW. Como tal deve ser tratada de forma diferenciada em relação às outras dimensões. Usualmente está presente em todo Data Mart, pois o DW é histórico.
- Costuma ser complexa no mundo real:
 - Dia, Mês, Trimestre, Semestre, Ano
 - Dia Acumulado no Mês, no Ano
 - Período Fiscal, Semana de Cinco Dias
 - Feriados, Fim de semana
- Qual a granularidade ideal? É claro, depende do projeto
 - Com granularidade diária, podemos organizar os dados por dias, meses, anos, por períodos fiscais (artificiais) da empresa, etc. Essa modelagem, é mais flexível a mudanças nos requisitos do negócio.
- Diferente das outras dimensões, a tabela pode ser carregada antecipadamente, de uma só vez e não requer fonte de dados
 - Exemplo: 5 anos passados + 5 anos futuros = 10 anos = 3.650 dias (linhas na tabela)

Date Dimension
Date Key (PK)
Date
Full Date Description
Day of Week
Day Number in Epoch
Week Number in Epoch
Month Number in Epoch
Day Number in Calendar Month
Day Number in Calendar Year
Day Number in Fiscal Month
Day Number in Fiscal Year
Last Day in Week Indicator
Last Day in Month Indicator
Calendar Week Ending Date
Calendar Week Number in Year
Calendar Month Name
Calendar Month Number in Year
Calendar Year-Month (YYYY-MM)
Calendar Quarter
Calendar Year-Quarter
Calendar Half Year
Calendar Year
Fiscal Week
Fiscal Week Number in Year
Fiscal Month
Fiscal Month Number in Year
Fiscal Year-Month
Fiscal Quarter
Fiscal Year-Quarter
Fiscal Half Year
Fiscal Year
Holiday Indicator
Weekday Indicator
Selling Season
Major Event
SQL Date Stamp
... and more

POS Retail Sales Transaction Fact
Date Key (FK)
Product Key (FK)
Store Key (FK)
Promotion Key (FK)
POS Transaction Number
Sales Quantity
Sales Dollar Amount
Cost Dollar Amount
Gross Profit Dollar Amount



Dimensão Tempo (Data)

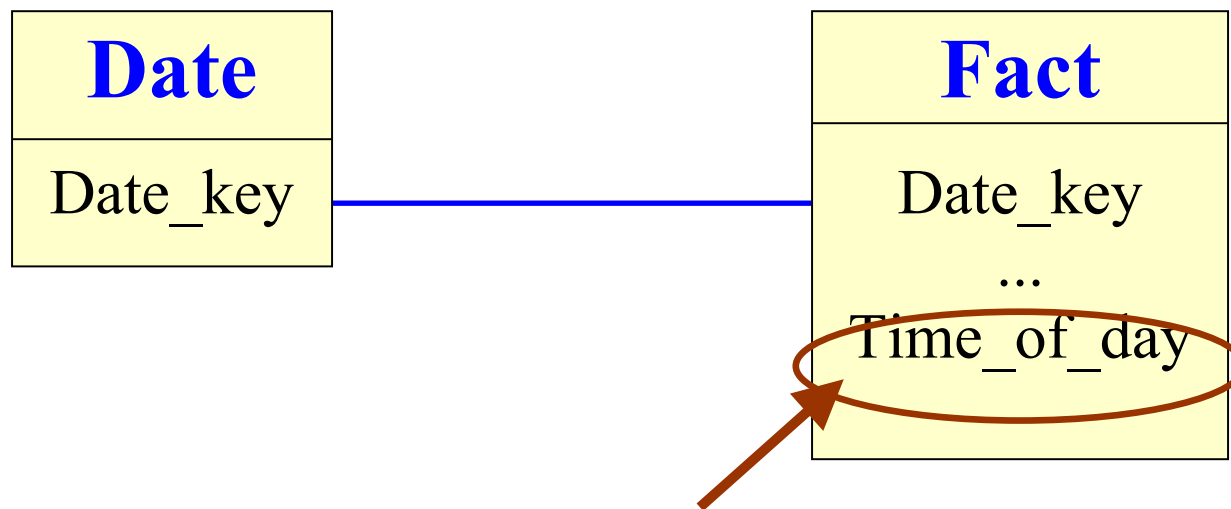
Date Key	Date	Full Date Description	Day of Week	Calendar Month	Calendar Year	Fiscal Year-Month	Holiday Indicator	Weekday Indicator
1	01/01/2002	January 1, 2002	Tuesday	January	2002	F2002-01	Holiday	Weekday
2	01/02/2002	January 2, 2002	Wednesday	January	2002	F2002-01	Non-Holiday	Weekday
3	01/03/2002	January 3, 2002	Thursday	January	2002	F2002-01	Non-Holiday	Weekday
4	01/04/2002	January 4, 2002	Friday	January	2002	F2002-01	Non-Holiday	Weekday
5	01/05/2002	January 5, 2002	Saturday	January	2002	F2002-01	Non-Holiday	Weekend
6	01/06/2002	January 6, 2002	Sunday	January	2002	F2002-01	Non-Holiday	Weekend
7	01/07/2002	January 7, 2002	Monday	January	2002	F2002-01	Non-Holiday	Weekday
8	01/08/2002	January 8, 2002	Tuesday	January	2002	F2002-01	Non-Holiday	Weekday

Tipo de dados SQL (Date, Time) não suportam essa riqueza de descrições, daí a necessidade de uma dimensão Data explícita.

Dimensão tempo: Horas, Minutos, Segundos

Várias soluções são possíveis, graças à extensibilidade do modelo dimensional.

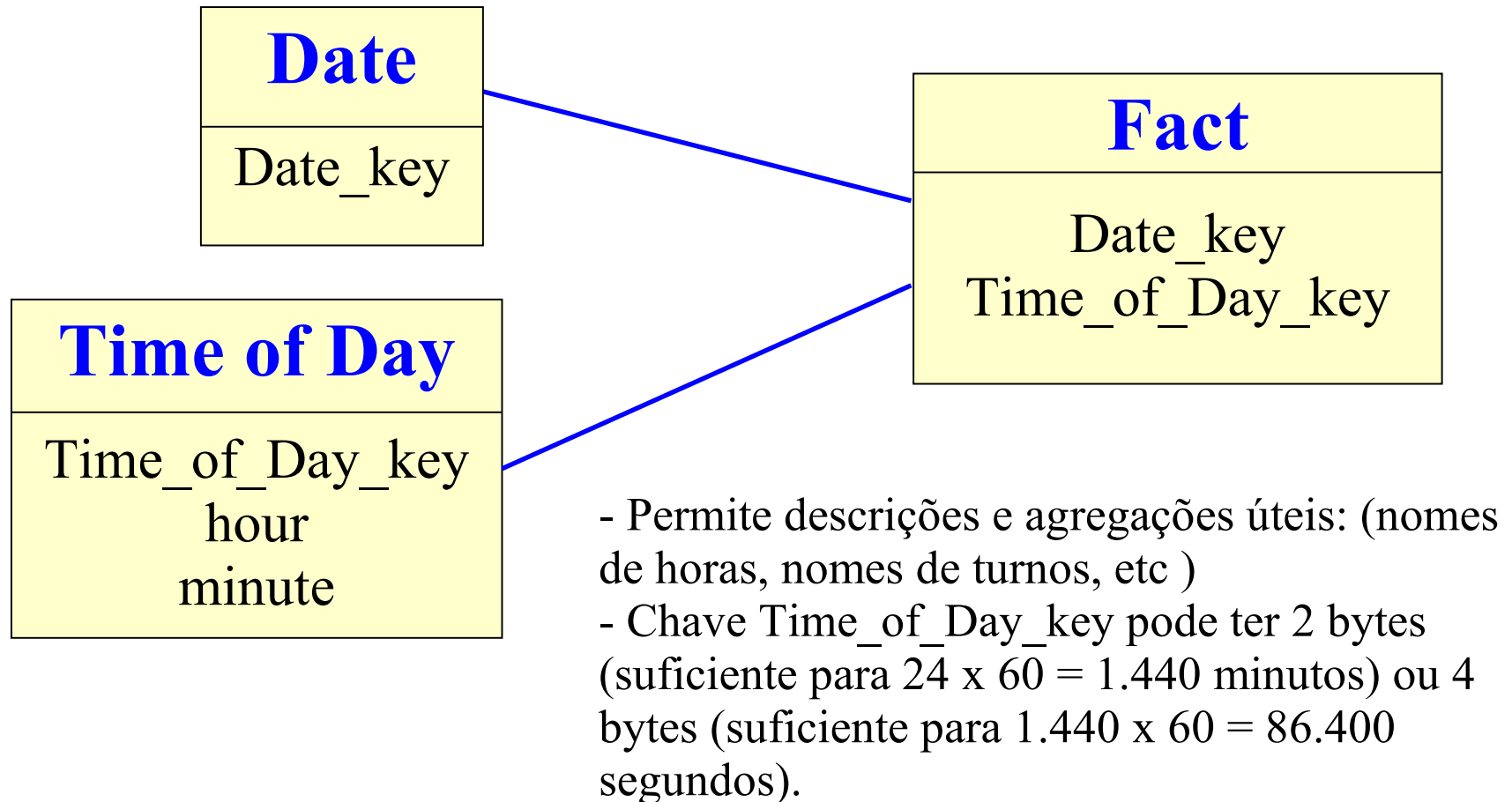
1ª Alternativa: Colocar a “hora do dia” na Tabela de Fatos



- Pode ser usado quando não há descrições adicionais sobre a hora do dia.
- Pode sobrecarregar a tabela de fatos (tipo Timestamp requer 8 bytes)
8 bytes x bilhões de linhas na tabela de fatos ...

Dimensão tempo: Horas, Minutos, Segundos

2ª Alternativa: Criar uma Dimensão Hora do Dia
(24 h X 60 min = 1440 valores)



Dimensão tempo: Horas, Minutos, Segundos

3ª Alternativa : Hora, minuto na mesma tabela de dimensão que as datas

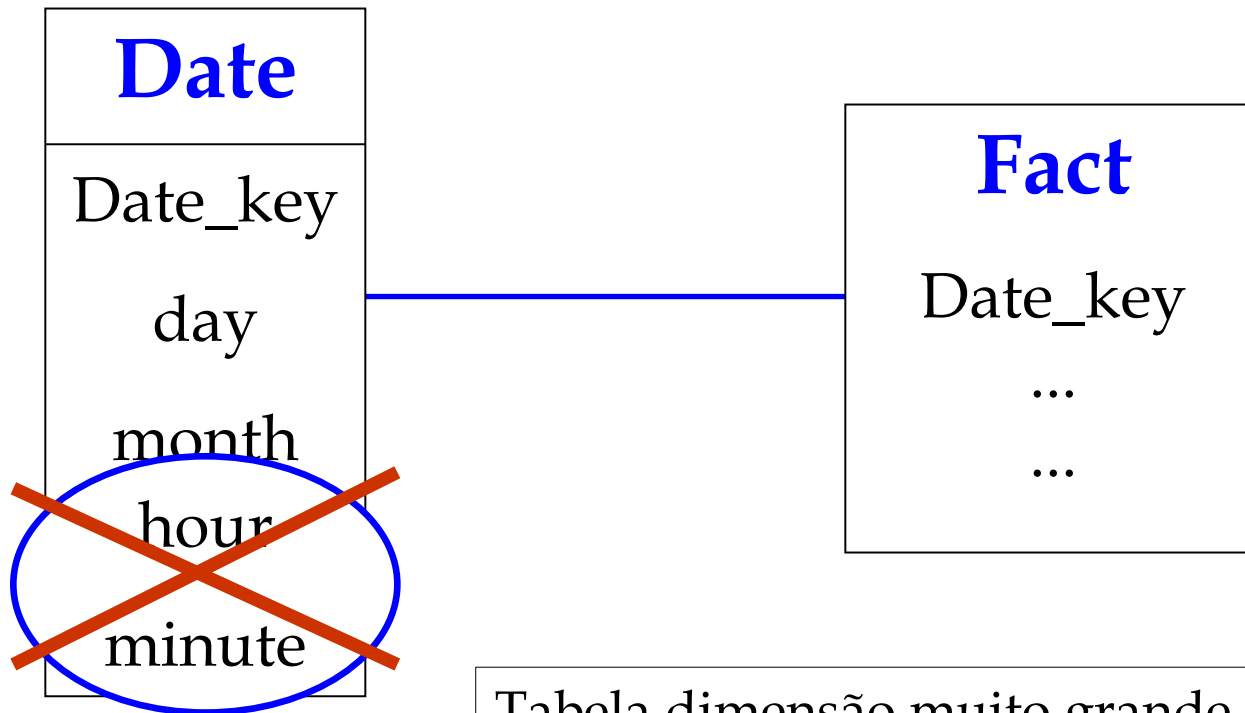
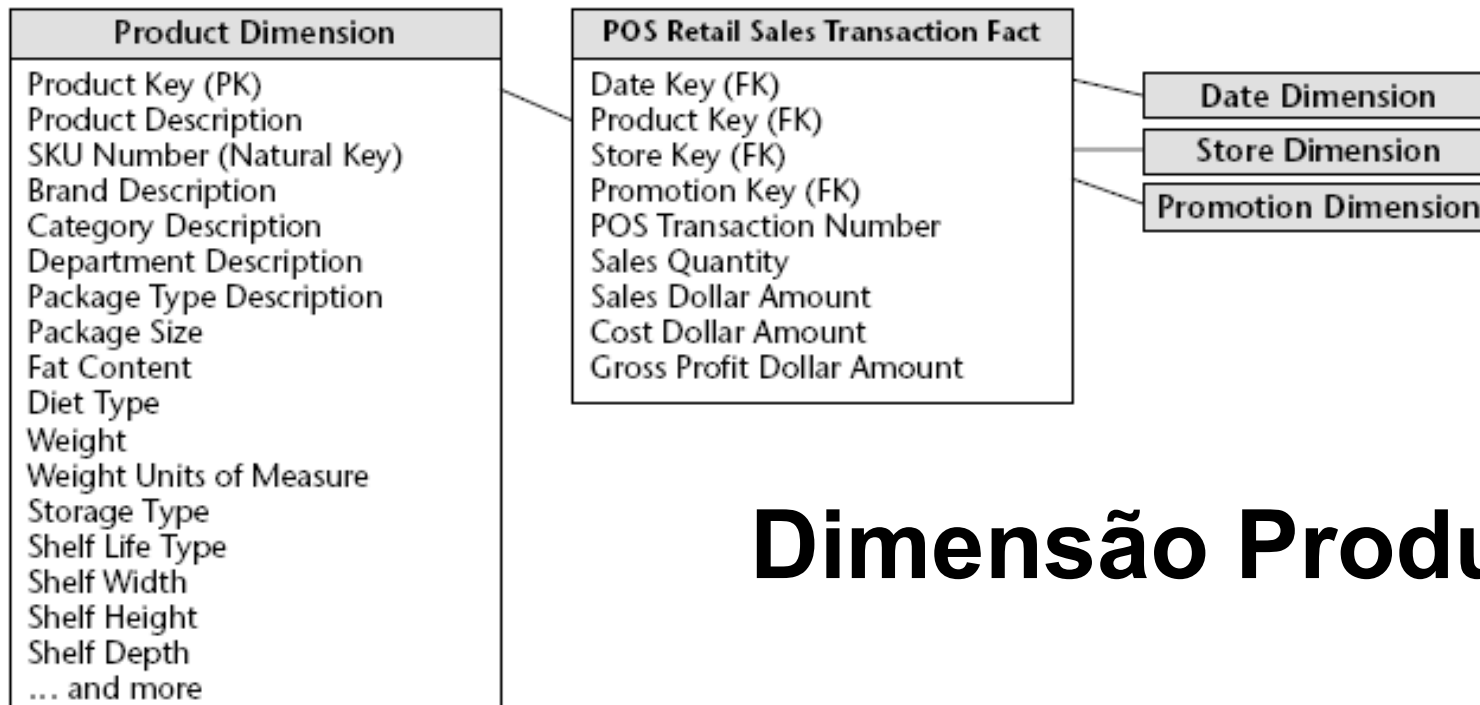


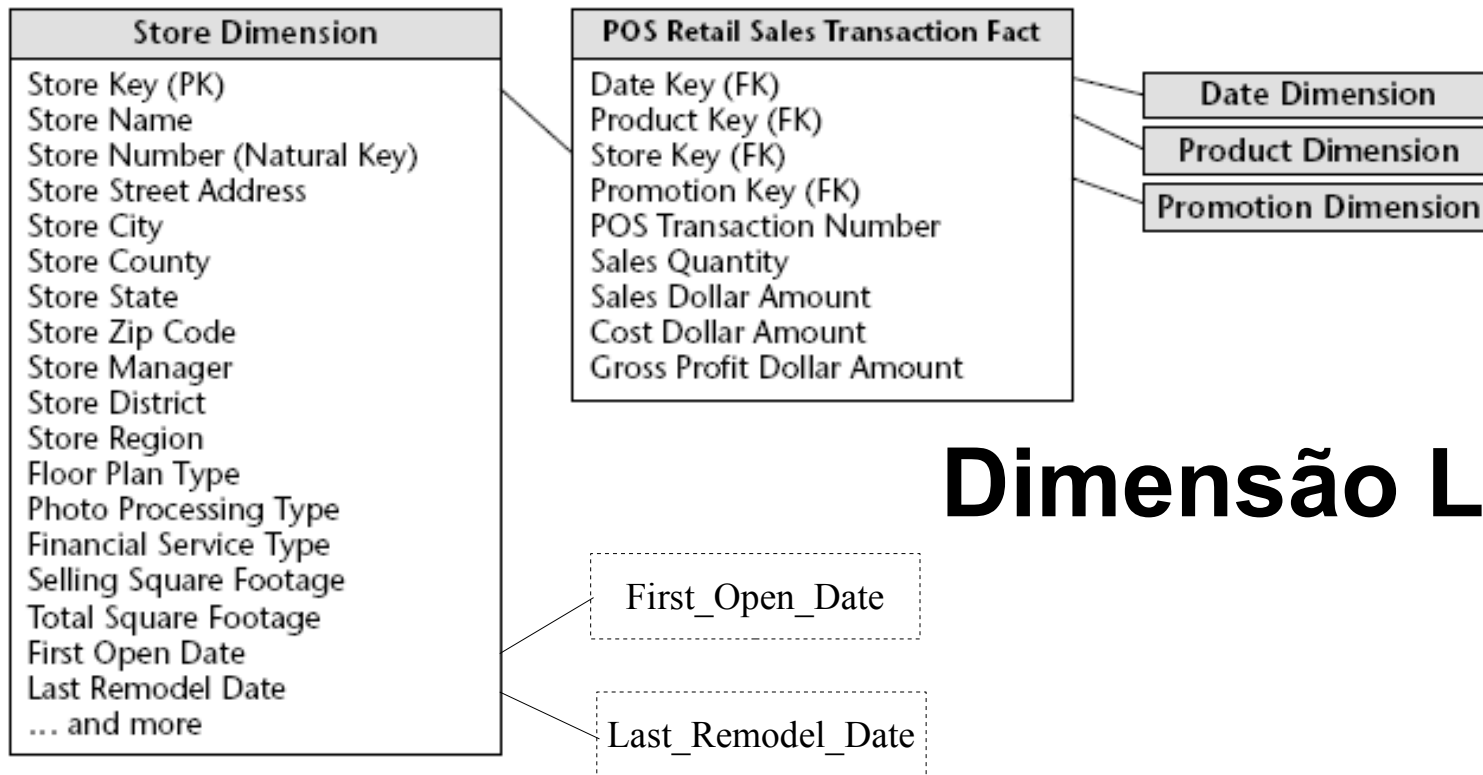
Tabela dimensão muito grande
10 anos = 3.650×1.440 minutos = 5.256.000
linhas (525.600 linhas cada ano adicional)



Dimensão Produto

Product Key	Product Description	Brand Description	Category Description	Department Description	Fat Content
1	Baked Well Light Sourdough Fresh Bread	Baked Well	Bread	Bakery	Reduced Fat
2	Fluffy Sliced Whole Wheat	Fluffy	Bread	Bakery	Regular Fat
3	Fluffy Light Sliced Whole Wheat	Fluffy	Bread	Bakery	Reduced Fat
4	Fat Free Mini Cinnamon Rolls	Light	Sweeten Bread	Bakery	Non-Fat
5	Diet Lovers Vanilla 2 Gallon	Coldpack	Frozen Desserts	Frozen Foods	Non-Fat
6	Light and Creamy Butter Pecan 1 Pint	Freshlike	Frozen Desserts	Frozen Foods	Reduced Fat
7	Chocolate Lovers 1/2 Gallon	Frigid	Frozen Desserts	Frozen Foods	Regular Fat
8	Strawberry Ice Creamy 1 Pint	Icy	Frozen Desserts	Frozen Foods	Regular Fat
9	Icy Ice Cream Sandwiches	Icy	Frozen Desserts	Frozen Foods	Regular Fat

Redundância à custa da 3FN vale a pena, pois as tabelas de dimensões são pequenas em relação às tabelas de fatos.

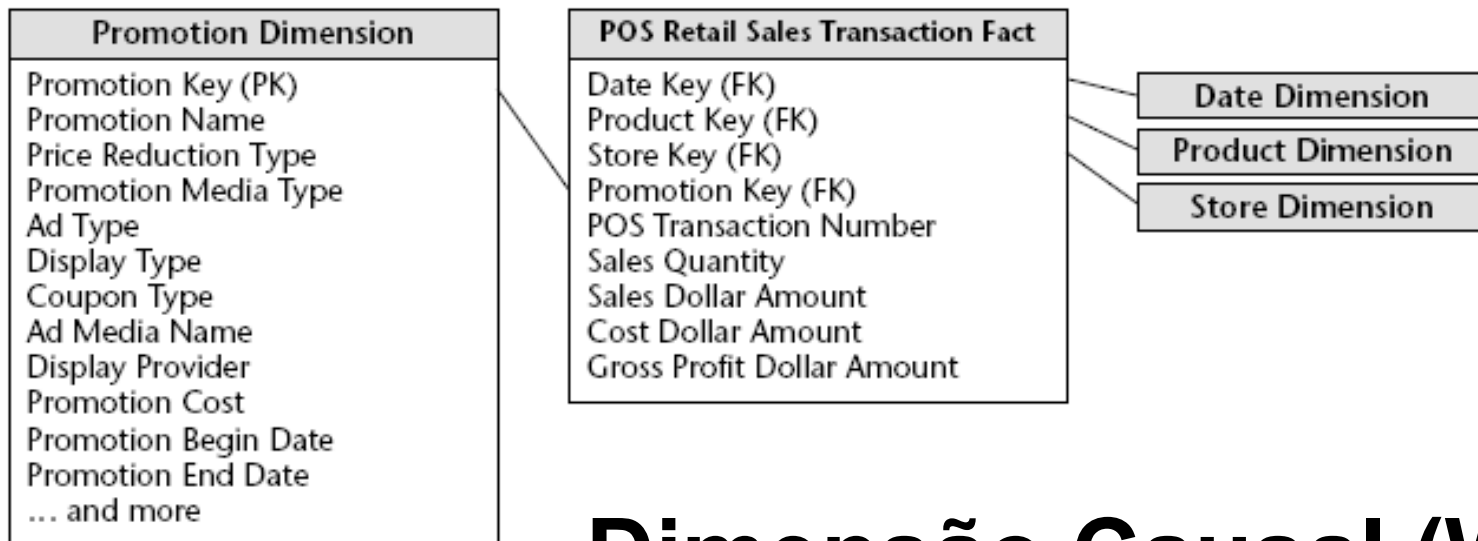


Dimensão Loja

Note os atributos First Open Date e Last Remodel Date, são DATAS. São chaves de junção com cópias da tabela de dimensão Date, declaradas como visões SQL, por exemplo

```
CREATE VIEW First_Open_Date (FO_day_number, FO_month, ...)
AS SELECT day_number, month
FROM Date
```

Esse tipo de tabela virtual para relacionar dimensões é denominado “outrigger”.



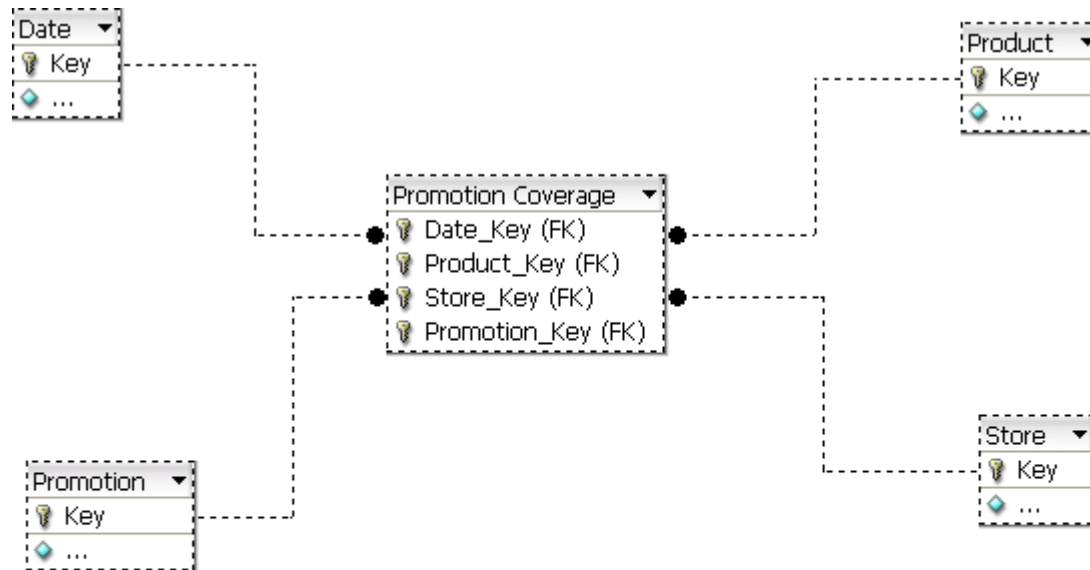
Dimensão Causal (Why) Promoção

- A dimensão Promoção do exemplo é, de fato uma COMBINAÇÃO DE DIMENSÕES causais (price reduction, ads, display, coupon) que poderiam estar em quatro tabelas separadas, com o mesmo efeito.
- No caso, estão combinadas numa única tabela de dimensão porque são altamente correlatas.
- Dimensões combinadas economizam espaço da tabela de Fatos, embora separadas pudessem ser mais bem entendidas e mais facilmente administradas.

Tabela de Fatos sem Fatos (Factless Fact Tables)

- Uma tabela de fatos que não tem fatos mas captura alguns relacionamentos muitos-para-muitos entre chaves de dimensões. Mais frequentemente usada para representar eventos ou prover informação de cobertura que não aparece em outras tabelas de fatos.
- A tabela de fatos Vendas com medidas não pode responder a consultas do tipo
 - Quais produtos estavam em promoção mas não venderam?
 - Por que não pode? Por que não deveria?
 - A solução é criar uma Tabela de Cobertura de promoção com as mesmas dimensões da tabela de Vendas (Data, Produto, Loja, Promoção).
 - Os produtos em promoção que não venderam será o conjunto diferença entre a cobertura e as vendas.

Tabela de Fatos sem Fatos Cobertura de Promoção

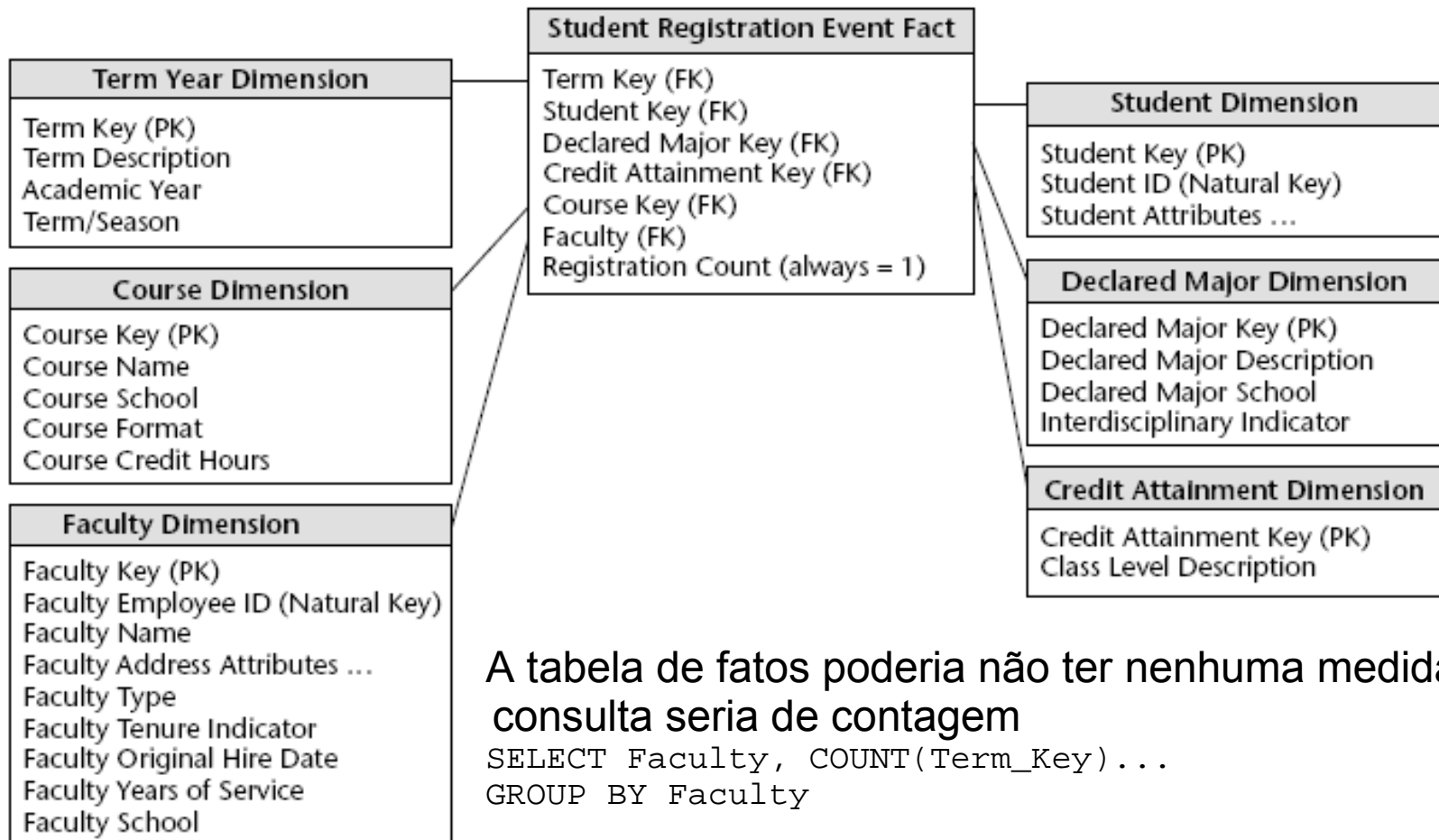


Uma tabela de fatos, tipicamente sem fatos, que registra todos os produtos que estão em promoção numa determinada loja, independentemente de ser vendidos ou não.

Consulta: Quais produtos estavam em promoção mas não venderam?

```
SELECT Product_Key, ... FROM Promotion_Coverage, ... WHERE ...  
MINUS  
SELECT Product_Key. ... FROM POS_Retail_Sales, ... WHERE ...
```

Tabela de Fatos sem Fatos - Eventos



A tabela de fatos poderia não ter nenhuma medida e a consulta seria de contagem

```
SELECT Faculty, COUNT(Term_Key)...  
GROUP BY Faculty
```

Ou poderia ter uma medida artificial `Registration_Count` apenas para tornar mais fácil a consulta

```
SELECT Faculty, SUM(Registration_Count)...  
GROUP BY Faculty
```

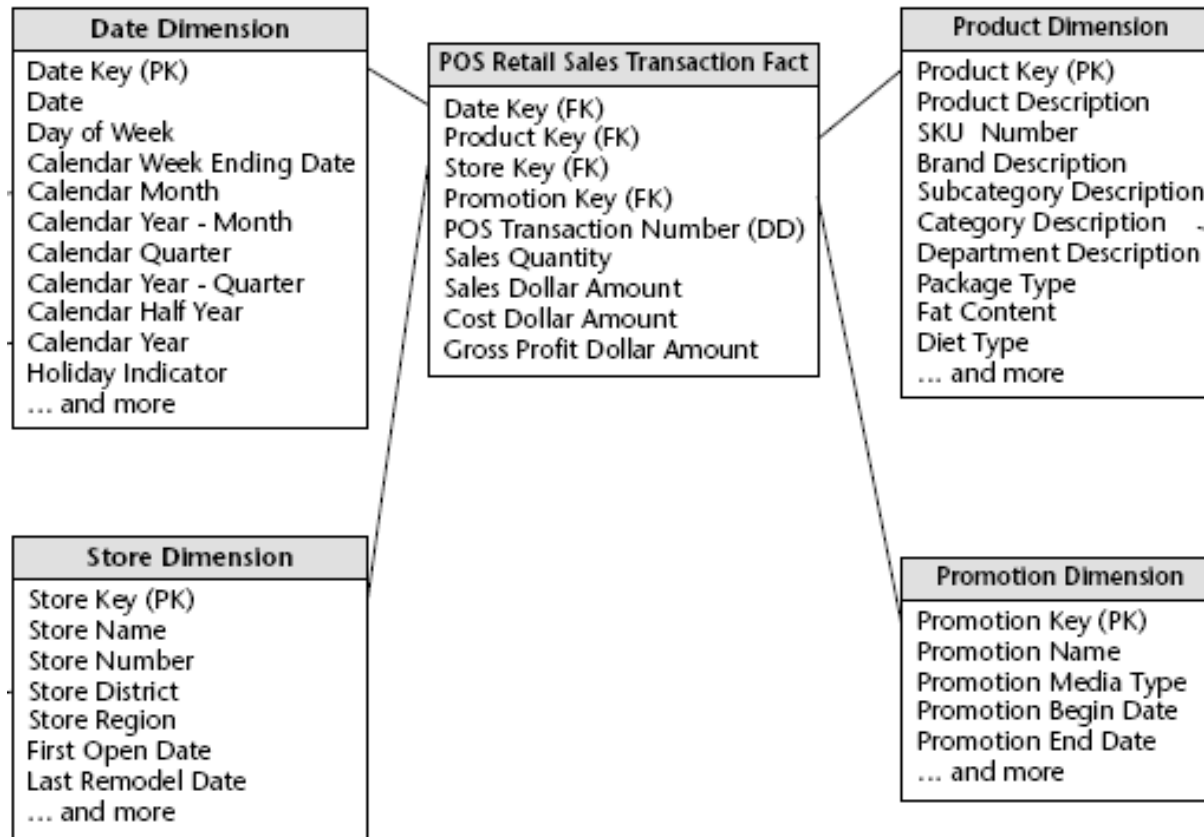
Dimensões sem Tabelas

Dimensões Degeneradas

- Chaves de dimensão na tabela de fatos sem tabelas de dimensão correspondentes.
- Uma chave de dimensão, como o número de uma transação, número de fatura, tiquete, nota fiscal, pedido ou ordem de compra, que não tenha nenhum atributo portanto não se junta com uma tabela de dimensão.
- Esses documentos normalmente são compostos de itens, e se a granularidade da tabela de fatos for item, o número do documento estará na tabela fato apenas para permitir o agrupamento dos itens por documento.

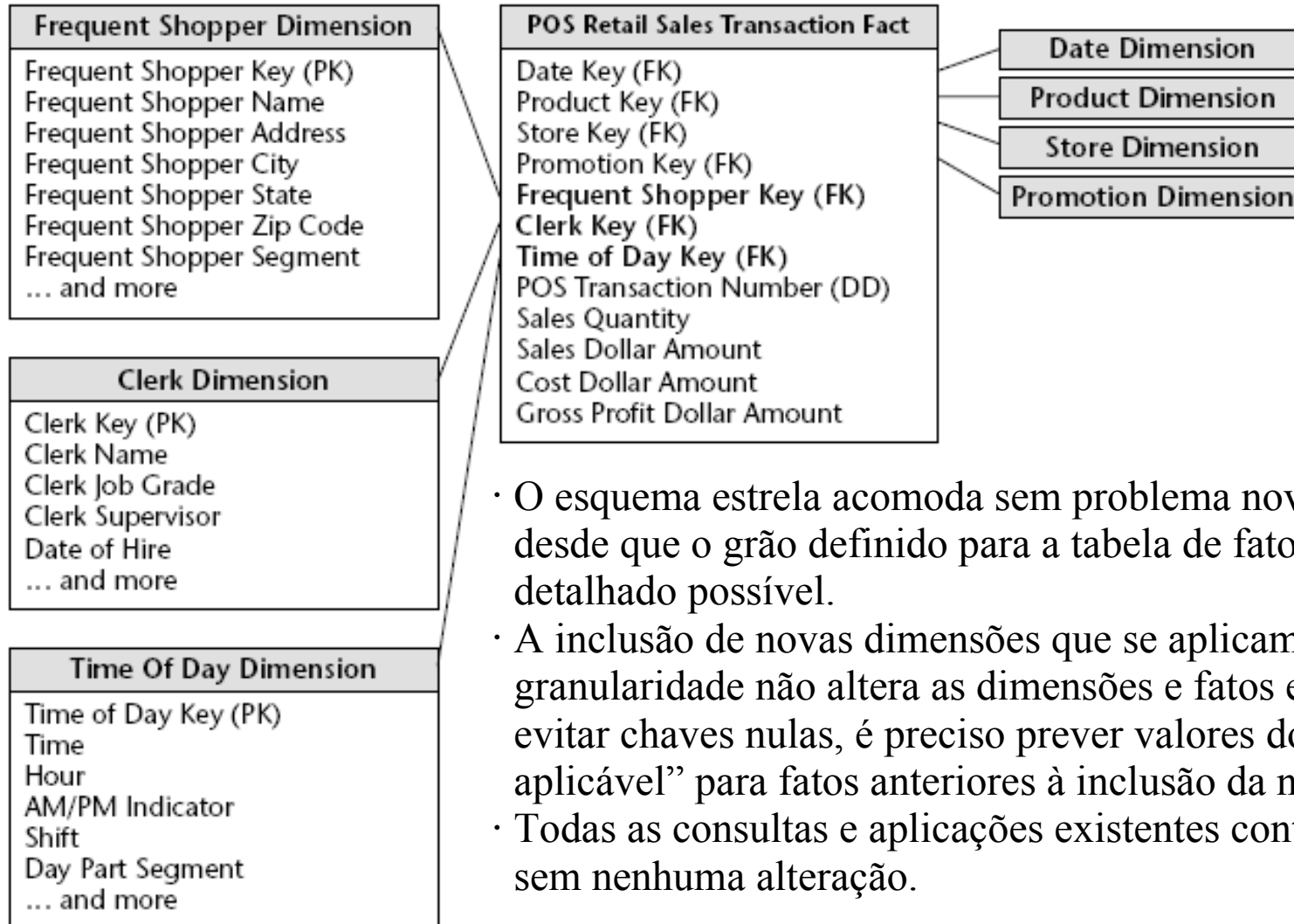
Dimensões sem Tabelas

Dimensões Degeneradas



POS Transaction Number é uma Dimensão Degenerada (DD)

Extensibilidade do Esquema Estrela



- O esquema estrela acomoda sem problema novas dimensões desde que o grão definido para a tabela de fatos seja o mais detalhado possível.
- A inclusão de novas dimensões que se aplicam a esse nível de granularidade não altera as dimensões e fatos existentes. Para evitar chaves nulas, é preciso prever valores do tipo “Não aplicável” para fatos anteriores à inclusão da nova dimensão.
- Todas as consultas e aplicações existentes continuam a rodar sem nenhuma alteração.

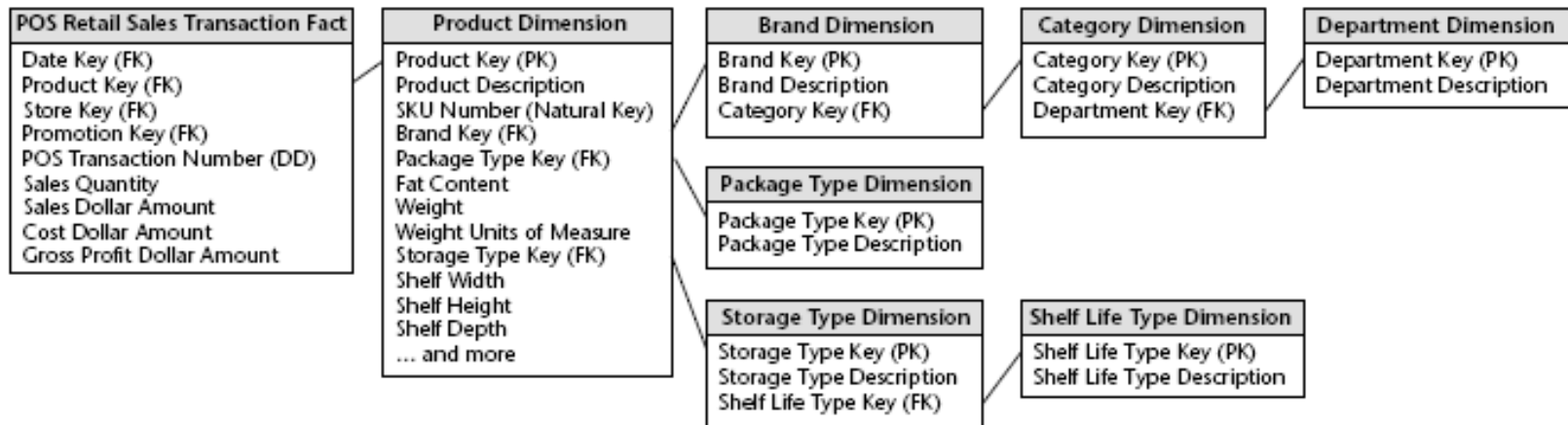
Extensibilidade do Esquema Estrela

Modificações absorvidas naturalmente pelo esquema estrela, devido a mudança nas fontes ou por decisão de modelagem, sem impacto nas aplicações existentes

- **Novos atributos de dimensões**
- **Novas dimensões**
- **Novos fatos medidos (na mesma tabela de fatos ou em nova tabela)**
- **Dimensões mais granulares**
- **Adição de uma fonte de dados nova envolvendo dimensões existentes assim como novas dimensões não previstas**

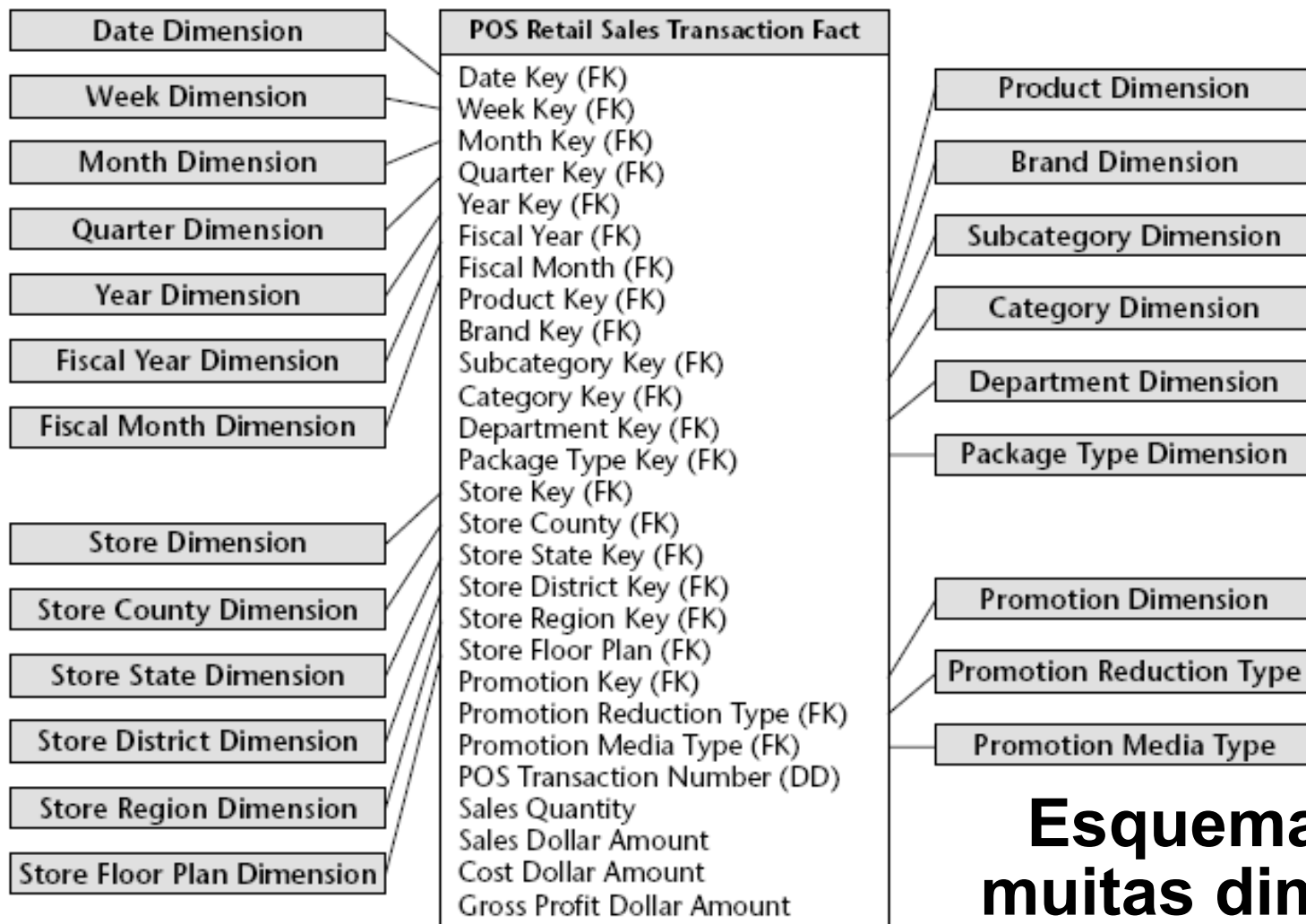
A extensibilidade é possível graças à simetria do esquema estrela, contanto que o grão inicial escolhido seja o mais detalhado possível pelos sistemas transacionais.

Esquema Dimensional Snow Flake



Embora aceitável, a normalização de dimensões não é recomendável por razões de desempenho e facilidade de uso

- **A quantidade de tabelas torna a apresentação do modelo mais complexa.**
- **Otimizadores do SGBD têm mais dificuldade com esquema complexo.**
- **A economia de espaço em disco é insignificante em relação ao DW completo.**
- **Snowflaking diminui a habilidade de usuários de navegar na dimensão.**
- **Snowflaking impede o uso de índices tipo Bit Map, que são usados por SGBD para indexar campos com baixa cardinalidade.**



Esquemas com muitas dimensões (Centípedo)

Um número de dimensões muito grande (25+) é um sinal de que muitas dimensões não são completamente independentes e deveriam ser combinadas numa única. É um erro em modelagem dimensional representar elementos de uma hierarquia como dimensões separadas.

Campos Chaves de Tabela de Dimensões

- **Regra básica: uso de surrogates ou chaves artificiais.**
 - Ajudam a manter a estabilidade, através da neutralidade.
 - Evitam manutenção custosa de tabelas, especialmente das tabelas fatos.
 - Chaves naturais podem ter problemas de unicidade, ausência, tamanhos exagerados.
 - Chaves artificiais podem ser especificadas como inteiros de 4 bytes, alcançando até 2^{32} , isto é, mais de 2 bilhões de ocorrências (inteiros positivos), o que é mais do que necessário para qualquer tabela dimensão.
 - Chaves artificiais ficam transparentes (invisíveis) para os usuários, servindo apenas como ligação entre dimensões e fatos.
 - Campos naturais não chave poderão ser indexados, tornando as consultas amistosas.
 - Se produzidas automaticamente, deve-se ter cuidado no processo de preparação (ETL), especialmente nos reprocessamentos.
 - A única desvantagem das chaves artificiais é que não faz sentido a tabela fato ser consultada diretamente, pois os campos descritivos de filtro estarão armazenados nas dimensões.
- **Every join between dimension and fact tables in the data warehouse should be based on meaningless integer surrogate keys. You should avoid using the natural operational production codes. None of the data warehouse keys should be smart, where you can tell something about the row just by looking at the key.**

Dinâmica das Dimensões

- Atualização das dimensões que mudam lentamente (Slowly Changing Dimensions)
 - Exemplos: Endereço de Cliente, Descrição de Produto.
- Várias alternativas
 - Tipo 1: Atualizar por cima do valor antigo
 - » É simples mas não preserva histórico.
 - Tipo 2: Adicionar uma nova linha com o novo valor do atributo atualizado, mantendo os demais.
 - » A nova linha particiona o histórico na tabela fato.
 - » É a técnica predominante para dimensões que mudam lentamente (slowly changing dimensions).
 - Tipo 3: Adicionar uma nova coluna, preservando o valor anterior e inserindo o novo valor na nova coluna.
 - » Permite a manutenção de duas visões simultâneas do histórico, mas dá margem a muitos valores nulos quando as mudanças são lentas.
 - Soluções híbridas, com múltiplas versões (linhas) combinadas ou não com coluna de valor anterior.
 - » Mais flexíveis e completas, porém mais complexas.

Slowly Changing Dimensions

Exemplos Tipo 1, Tipo 2, Tipo 3

Linha original

Product Key	Product Description	Department	SKU Number (Natural Key)
12345	IntelliKidz 1.0	Education	ABC922-Z

Mudança: O produto IntelliKidz 1.0 muda de departamento.

SCD Tipo 1

Product Key	Product Description	Department	SKU Number (Natural Key)
12345	IntelliKidz 1.0	Strategy	ABC922-Z

SCD Tipo 2

Product Key	Product Description	Department	SKU Number (Natural Key)
12345	IntelliKidz 1.0	Education	ABC922-Z
25984	IntelliKidz 1.0	Strategy	ABC922-Z

SCD Tipo 3

Product Key	Product Description	Department	Prior Department	SKU Number (Natural Key)
12345	IntelliKidz 1.0	Strategy	Education	ABC922-Z

SCD: Exemplo Tipo Híbrido (também chamado tipo 6 = 3+2+1)

Linha original

Product Key	Product Description	Current Department	Historical Department	SKU Number (Natural Key)
12345	IntelliKidz 1.0	Education	Education	ABC922-Z

Requisito: Preservar histórico e ao mesmo tempo suportar consultas a dados históricos de acordo com valores atuais.

Primeira mudança

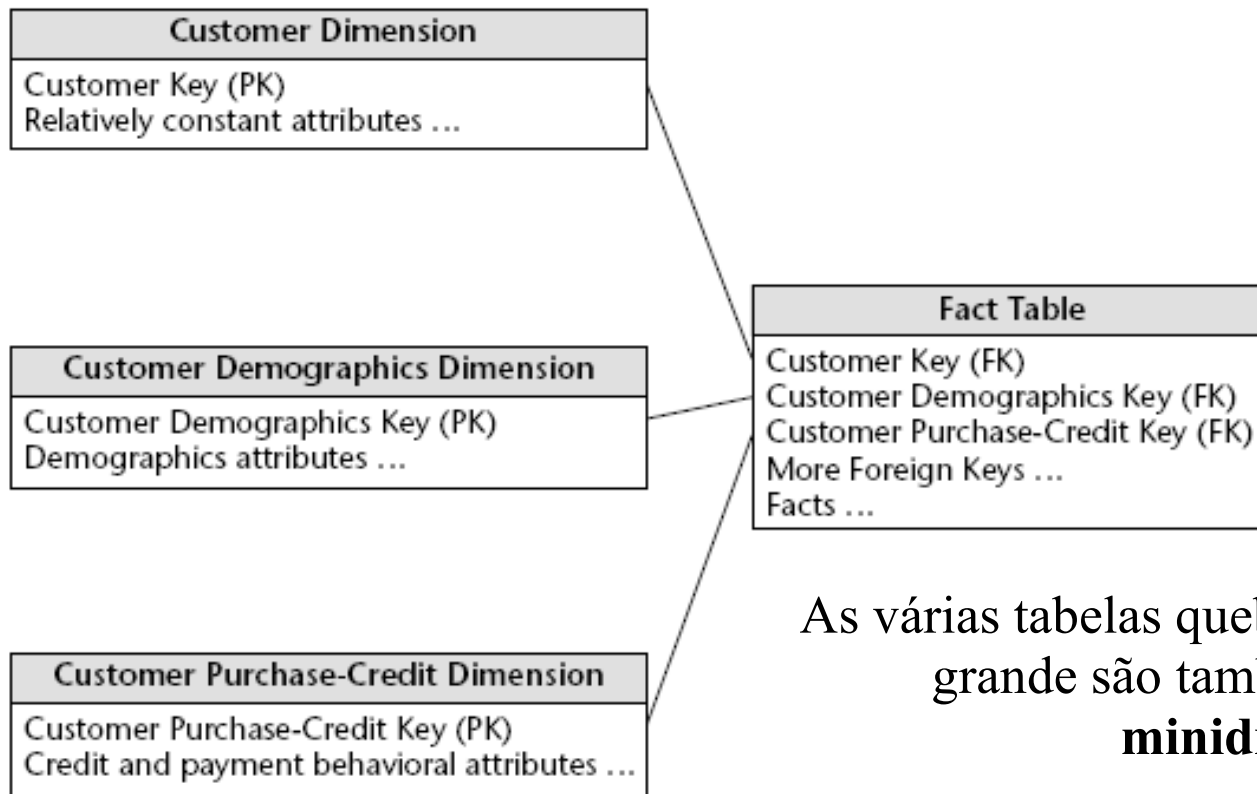
Product Key	Product Description	Current Department	Historical Department	SKU Number (Natural Key)
12345	IntelliKidz 1.0	Strategy	Education	ABC922-Z
25984	IntelliKidz 1.0	Strategy	Strategy	ABC922-Z

Segunda mudança

Product Key	Product Description	Current Department	Historical Department	SKU Number (Natural Key)
12345	IntelliKidz 1.0	Critical Thinking	Education	ABC922-Z
25984	IntelliKidz 1.0	Critical Thinking	Strategy	ABC922-Z
31726	IntelliKidz 1.0	Critical Thinking	Critical Thinking	ABC922-Z

Dimensões com grande volume e alta volatilidade também chamadas de Rapidly Changing Monster Dimensions

- Solução para dimensões grandes com mudanças frequentes (por exemplo, alguns atributos mudam mensalmente)
 - » Particionamento da dimensão em tabelas diferentes, separando-se dados estáticos de dados voláteis.
 - Dimensões são relacionadas entre si e ambas relacionadas com a tabela fato

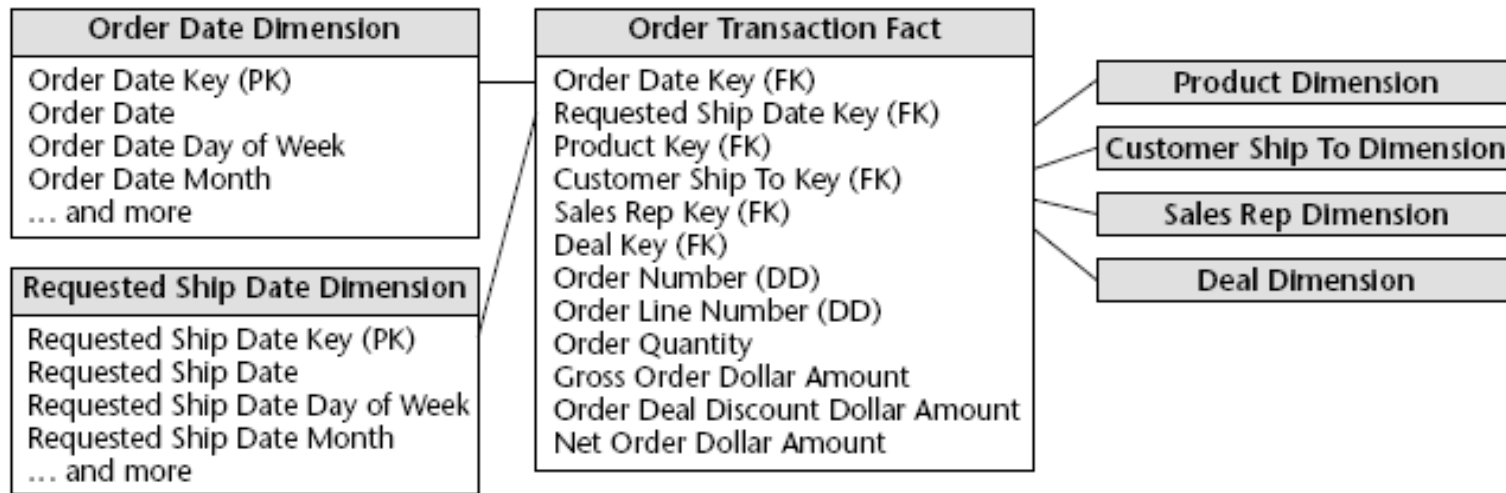


As várias tabelas quebradas de uma dimensão grande são também chamadas de **minidimensões**

Dimensões com vários Papéis

Role Playing Dimensions

A situação onde uma mesma dimensão aparece várias vezes na mesma tabela de fatos. Cada um dos papéis da dimensão é representado por uma tabela lógica separada com nomes de coluna únicos através de visões.



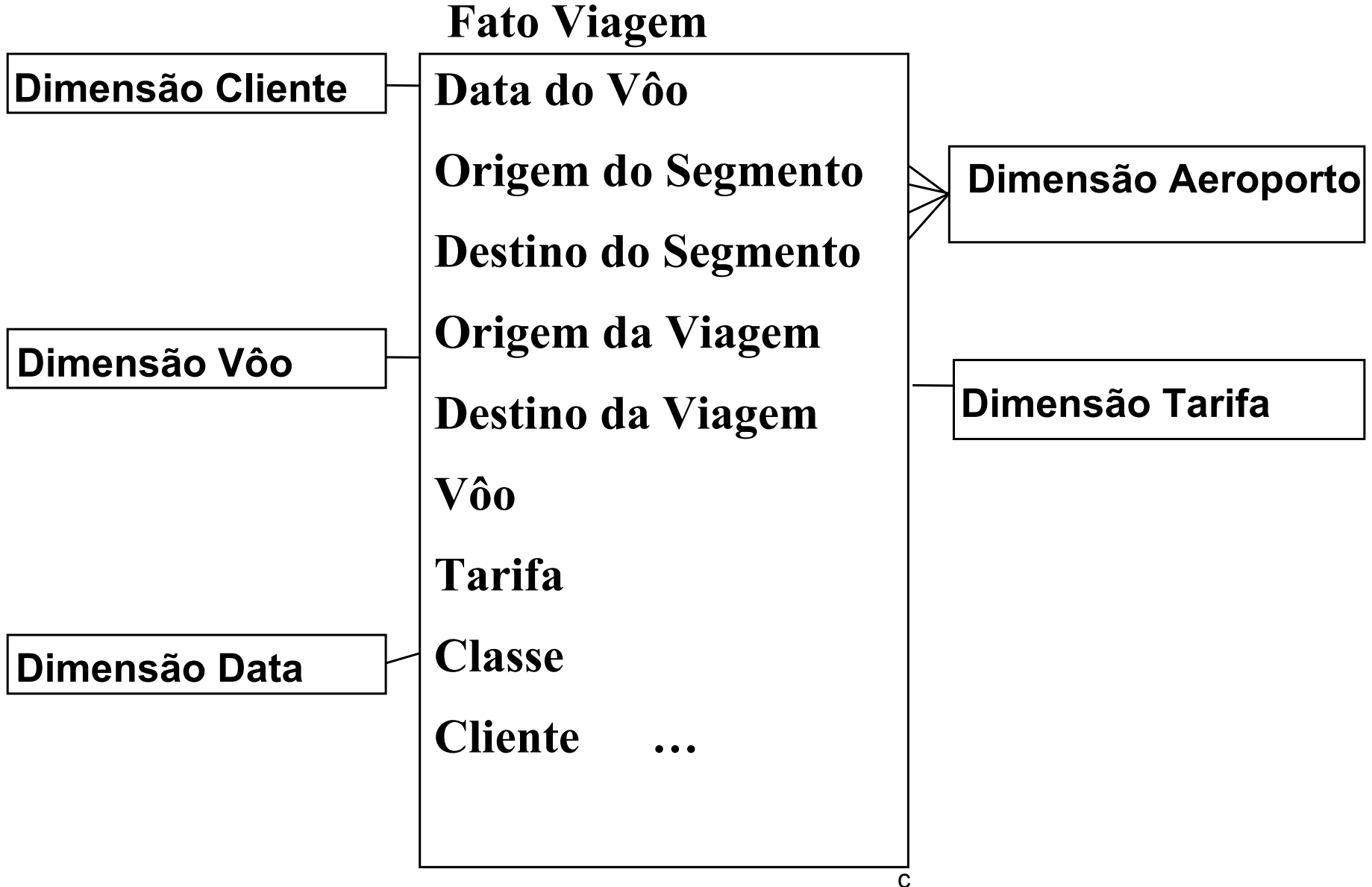
```
CREATE VIEW order_date (order_date_key, order_day_of_week,  
order_month...)
```

```
AS SELECT date_key, day_of_week, month, . . . FROM Date
```

```
CREATE VIEW req_ship_date (req_ship_date_key, req_ship_day_of_week,  
req_ship_month ...)
```

```
AS SELECT date_key, day_of_week, month, . . . FROM Date
```

Outros exemplos de Dimensões com papéis



Mais de uma dimensão com vários papéis

Dimensão Data

Dimensão Provedor

Dimensão Localização

Tráfego Tarifado de Comutação

Data da Chamada

Data da Tarifação

Data do Faturamento

Data do Pagamento

Provedor do Sistema de Origem

Provedor da Comutação Local

Provedor dos Interurbanos

Provedor do Serviço de Valor Agregado

Parte que Ligou

Parte que Recebeu a Ligação

Comutação Anterior

Comutação Subsequente

Outros Tipos Especiais de Dimensão

- **Dimensão lixo ou sucata (junk dimension)**
 - Uma dimensão abstrata com a decodificação de um grupo de flags e indicadores de baixa cardinalidade, portanto removendo os flags da tabela de fatos.
- **Minidimensões**
 - Subconjuntos de uma dimensão grande, como Cliente, que são quebrados em dimensões artificiais menores para controlar o crescimento explosivo de uma dimensão grande, com mudança rápida. Os atributos demográficos continuamente mutáveis de um cliente são frequentemente modelados como uma minidimensão separada.
- **Dimensões com “Outrigger”**
 - Solução normalizada (snow flake) para conjuntos de atributos de baixa cardinalidade em dimensões grandes, como Cliente. A economia de espaço vale a pena porque a dimensão é grande, e a carga de dados é separada do restante da dimensão porque os dados provêm de fontes externas diferentes.
- **Dimensões multivaloradas (tabela ponte)**
 - Normalmente, uma tabela de fatos possui conexões somente para dimensões representando um valor simples, como uma data ou produto. Mas ocasionalmente, é válido conectar um registro de fato a uma dimensão representando um número aberto de valores, como o número de diagnósticos simultâneos que um paciente pode ter num momento de um mesmo tratamento. Neste caso, dizemos que a tabela de fatos tem uma dimensão multivalorada. Tipicamente manipulada por uma tabela ponte (Bridge Table) também chamada Helper Table, Tabela Associativa).

Dimensão lixo ou sucata (junk dimension)

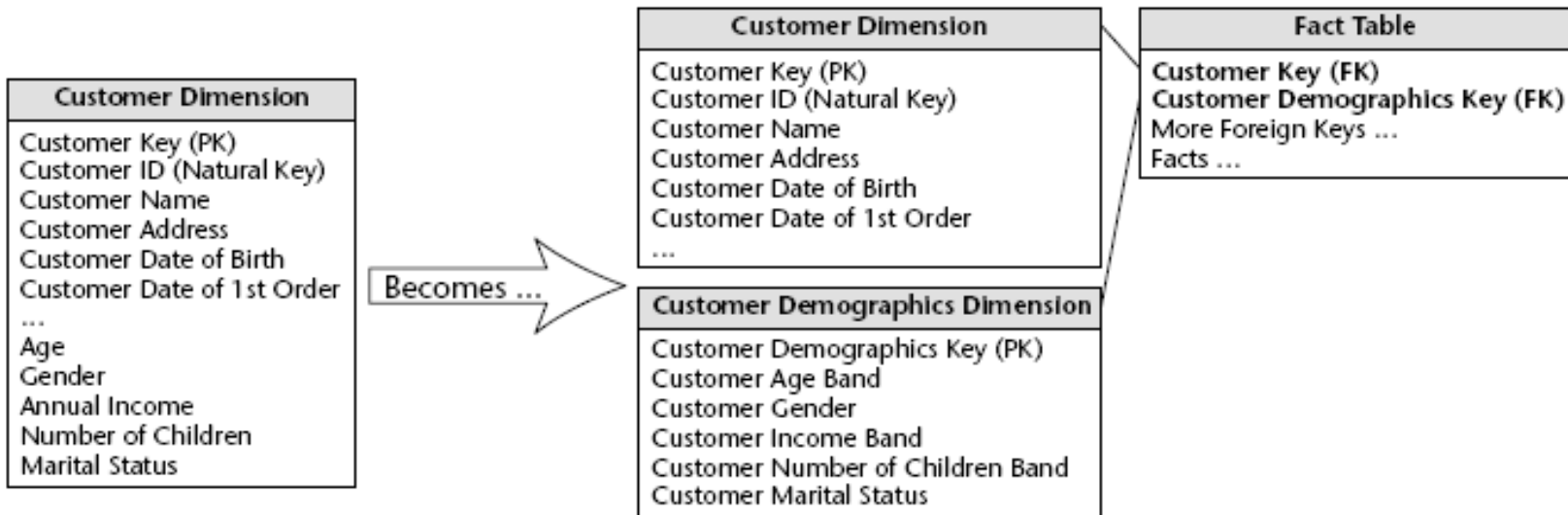
- Relacionadas com tabelas tipo código-descrição com baixa cardinalidade: Sexo, Estado Civil, Tags diversos, Textos descritivos, etc. São campos tipo miscelânea que não trazem muita correlação com os outros campos da tabela fato, mas são usados como filtro, daí serem dimensões.
- Podem ser usadas de forma combinada.
 - Exemplo: três tags binários $\rightarrow 2^3 = 8$ combinações possíveis
- Usado como artifício para diminuir a tabela de fatos. Exemplo:

Order Indicator Key	Payment Type Description	Payment Type Group	Inbound/ Outbound Order Indicator	Commission Credit Indicator	Order Type Indicator
1	Cash	Cash	Inbound	Commissionable	Regular
2	Cash	Cash	Inbound	Non-Commissionable	Display
3	Cash	Cash	Inbound	Non-Commissionable	Demonstration
4	Cash	Cash	Outbound	Commissionable	Regular
5	Cash	Cash	Outbound	Non-Commissionable	Display
6	Discover Card	Credit	Inbound	Commissionable	Regular
7	Discover Card	Credit	Inbound	Non-Commissionable	Display
8	Discover Card	Credit	Inbound	Non-Commissionable	Demonstration
9	Discover Card	Credit	Outbound	Commissionable	Regular
10	Discover Card	Credit	Outbound	Non-Commissionable	Display
11	MasterCard	Credit	Inbound	Commissionable	Regular
12	MasterCard	Credit	Inbound	Non-Commissionable	Display
13	MasterCard	Credit	Inbound	Non-Commissionable	Demonstration
14	MasterCard	Credit	Outbound	Commissionable	Regular

Figure 5.5 Sample rows of an order indicator junk dimension.

Minidimensões

A melhor abordagem para tratar atributos em dimensões muito grandes é quebrar em uma ou mais minidimensões, cada uma contendo atributos que tenham um número limitado de valores. Exemplo: dimensão Cliente com milhões de ocorrências.



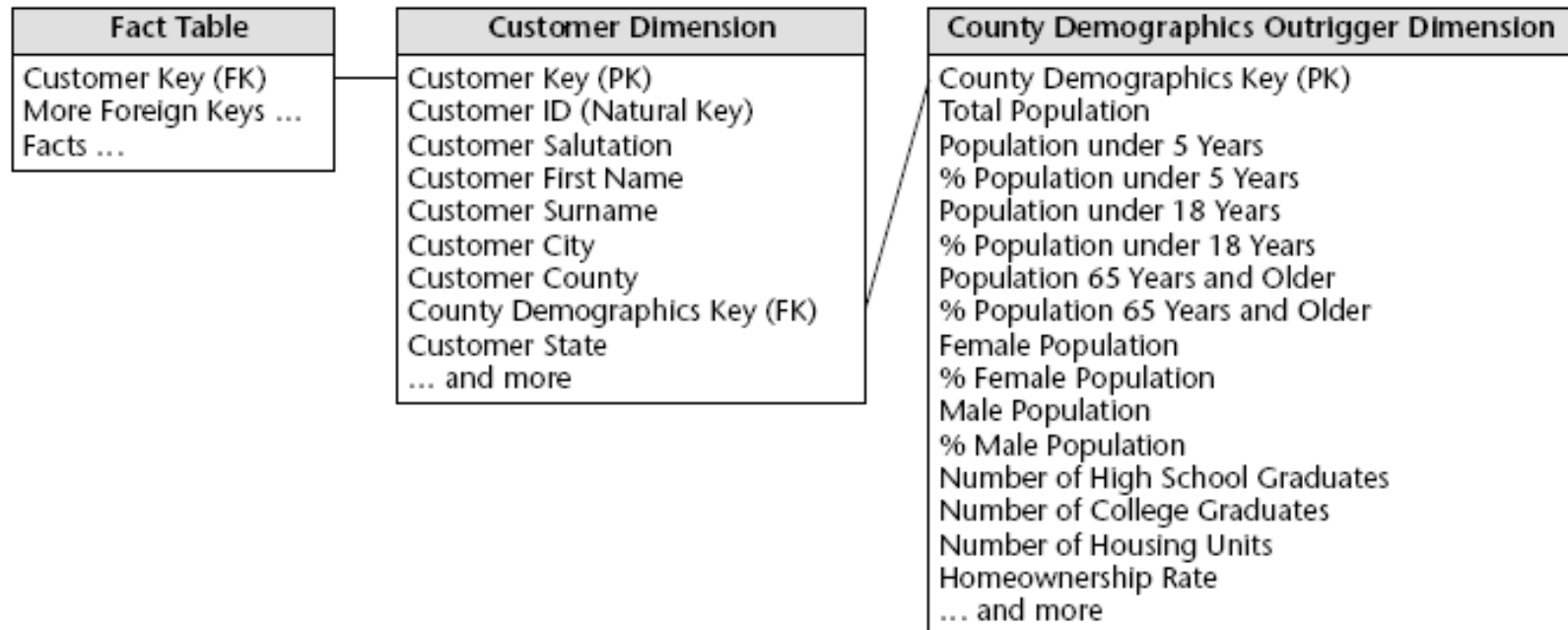
- Vide também o caso de dimensões com alta volatilidade (minidimensão com atributos que mudam rapidamente)
- Tamanho de cada minidimensão = Produto cartesiano da cardinalidade dos atributos da minidimensão

Exemplo acima: $10 \times 2 \times 10 \times 5 \times 5 = 5.000$ linhas

Dimensões com “Outrigger”

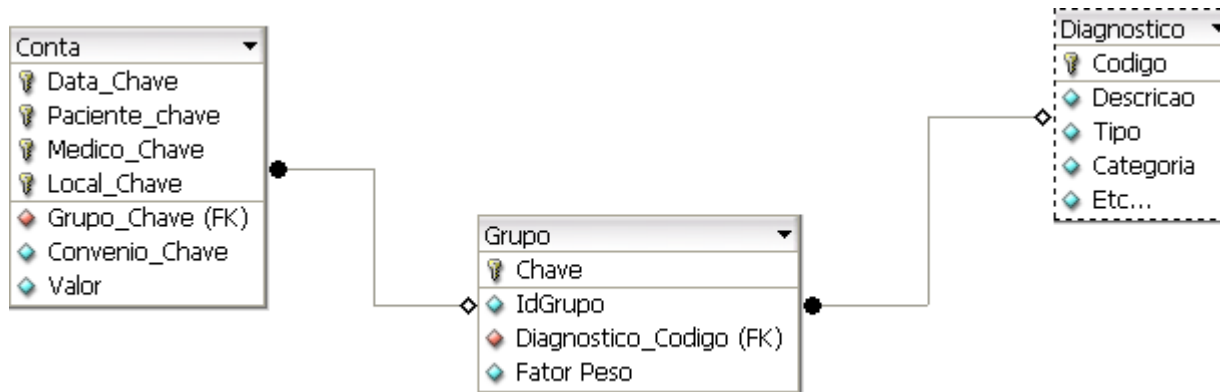
No exemplo, o “outrigger” agrupa atributos de baixa cardinalidade, que são mantidos em tabela separada da dimensão principal (Customer) para economia de espaço, e também porque a carga dessa tabela é feita com frequência diferente e a partir de fonte externa.

Note que se a solução fosse ligar o “outrigger” diretamente à tabela de Fatos, seria uma minidimensão. Seria possível? Vantagens e desvantagens?



Dimensões Multivaloradas (Tabela Ponte) (Bridge Table, Helper Table, Associative Table)

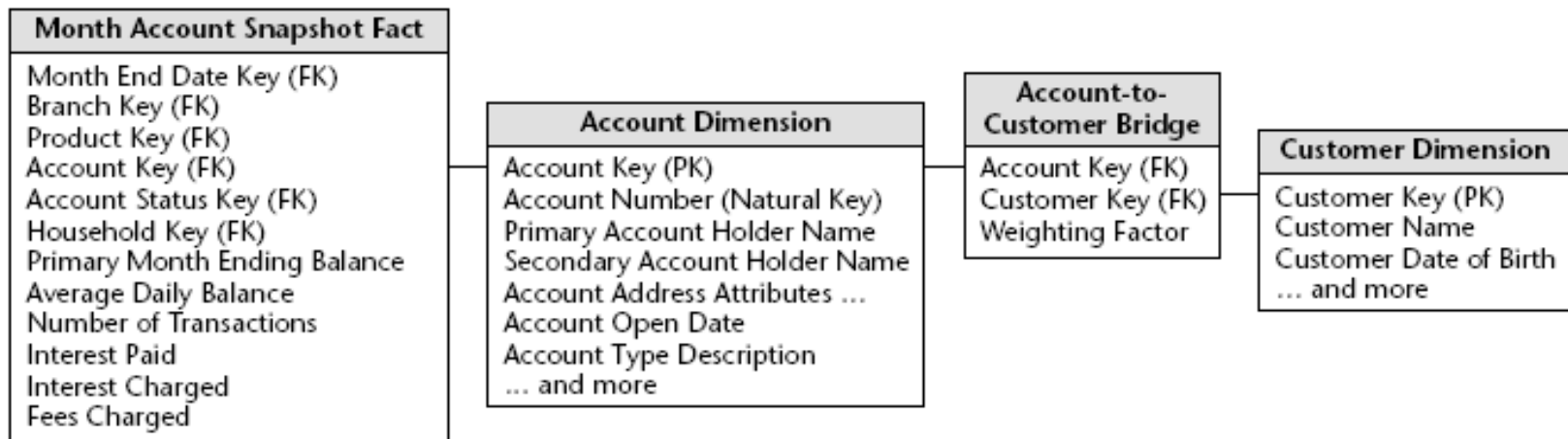
- Uma tabela com chave composta capturando um relacionamento muitos-para-muitos que não possa ser acomodado pela granularidade natural de uma tabela de fatos ou tabela de dimensão. Serve como uma ponte entre a tabela de fatos e a tabela de dimensão de forma a permitir dimensões multivaloradas.



- Outros exemplos de dimensões multivaloradas: titulares de conta bancária, códigos de classificações, etc

Tabela Ponte

Outro exemplo



- Tabela ponte conta-cliente para associar múltiplos clientes com fatos de contas.

Tópicos Especiais sobre Fatos

- **Fatos conformados**
 - Data marts de primeiro nível , data marts consolidados.
 - Unidades de medida
 - Bus Matrix de Implementação
- **Tipos clássicos de fatos**
 - Transações
 - Instantâneos Periódicos
 - Instantâneos Acumulados
- **Fatos agregados**

Fatos Conformados

- Estabelecer dimensões conformadas para amarrar os data marts representa 90% do esforço de arquitetura de projeto. O restante do esforço consiste em estabelecer definições de fatos conformados.
- Preços, custos, lucros, medidas de qualidade, medidas de satisfação do cliente e outros KPIs são fatos que devem ser conformados. Em geral, dados de fatos não são duplicados explicitamente em múltiplos data marts. Mas isso pode ocorrer em data marts de primeiro nível (originários de um sistema fonte primário de dados) e data marts consolidados (a partir de múltiplas fontes que podem referenciar mais de um processo de negócio).
- Se os fatos forem rotulados identicamente, precisam ser definidos no mesmo contexto dimensional e com as mesmas unidades de medida de data mart para data mart.
- Algumas vezes, um fato tem uma unidade de medida natural em uma tabela de fatos e outra unidade de medida em outra tabela de fatos. Ao invés de prover um fator de conversão numa tabela de dimensão, a abordagem correta é levar o fato com as duas unidades de medida para para facilitar os relatórios sem preocupação de conversão. Por exemplo, produtos medidos em caixas no depósito e em peças na loja.

Data Marts de Primeiro Nível e Data Marts Consolidados

	Time	Customer	Service	Area Category	Local Sec Provider	Calling Party	Called Party	Long Dist Provider	Internal Organization	Employee	Location	Equipment Type	Supplier	Item Supplier	Weather	Account Status
First-Level Marts:																
Customer Billing	X	X	X	X	X			X		X						X
Service Orders	X	X	X		X			X	X	X	X	X			X	X
Trouble Reports	X	X	X		X	X		X	X	X	X	X	X	X	X	X
Yellow Page Ads	X	X		X		X			X	X	X					X
Customer Inquiries	X	X	X	X	X	X		X	X	X	X				X	X
Promotions & Comm'n	X	X	X	X	X	X		X	X	X	X	X	X			X
Billing Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
Network Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
Customer Inventory	X	X	X	X	X			X	X		X	X	X	X		X
Network Inventory	X		X						X	X	X	X	X	X		
Real Estate	X								X	X	X	X				
Labor & Payroll	X								X	X	X					
Computer Charges	X	X	X		X			X	X	X	X	X	X	X		
Purchase Orders	X								X	X	X	X	X	X		
Supplier Deliveries	X								X	X	X	X	X	X		
Second-Level Marts:																
Combined Field Ops	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X
Cust Rein Mgmt	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Customer Profit	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

The Matrix Plan for the enterprise data warehouse of a large telecommunications company. First-level data marts are directly derived from production applications. Second-level data marts are developed later and represent combinations of first-level data marts.

Artigo “The Matrix”, Ralph Kimball, Intelligent Enterprise, Dezembro 1999.

http://www.intelligententerprise.com/db_area/archives/1999/990712/webhouse.jhtml

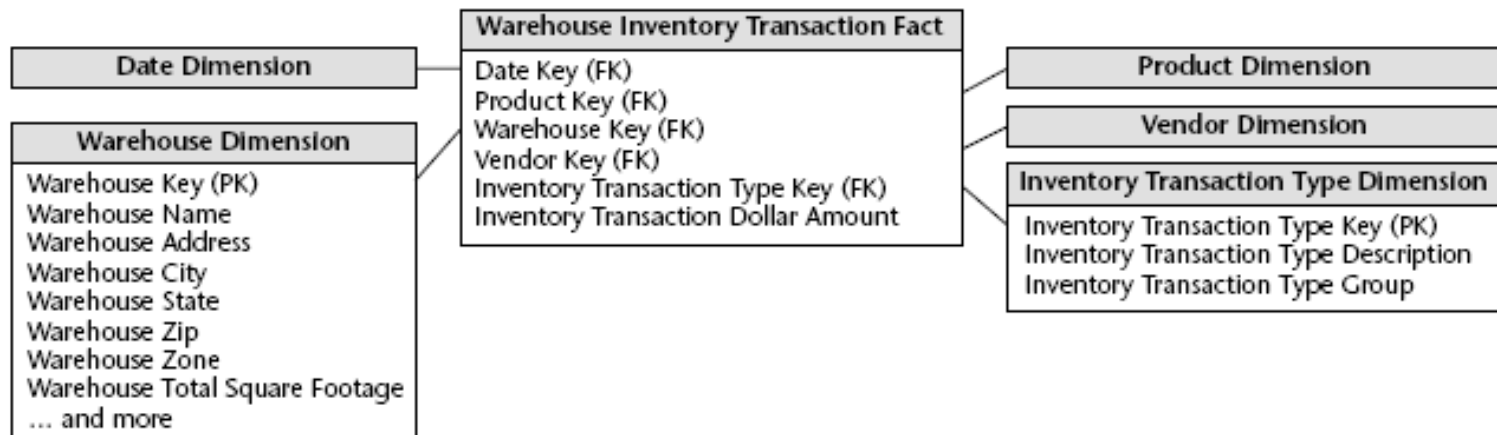
Bus Matrix de Implementação

- Exemplo em Negócio de Seguradora

Business Process	Fact Table	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 Bf	X	X	X	X	X			
	Auto Policy Transactions	1 row per auto policy transaction	Policy Transaction Amount	X Trxn Bf	X	X Auto	X Auto	X	X			
	Home Policy Transactions	1 row per home policy transaction	Policy Transaction Amount	X Trxn Bf	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, auto covered item, and auto 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, home covered item, and home coverage each month	Written Premium Revenue Amount, Earned Premium Revenue Amount	X	X	X Home	X Home	X Agent	X			
Claims Transactions	Claim Transactions	1 row for every claim transaction	Claim Transaction Amount	X Trxn Bf	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	

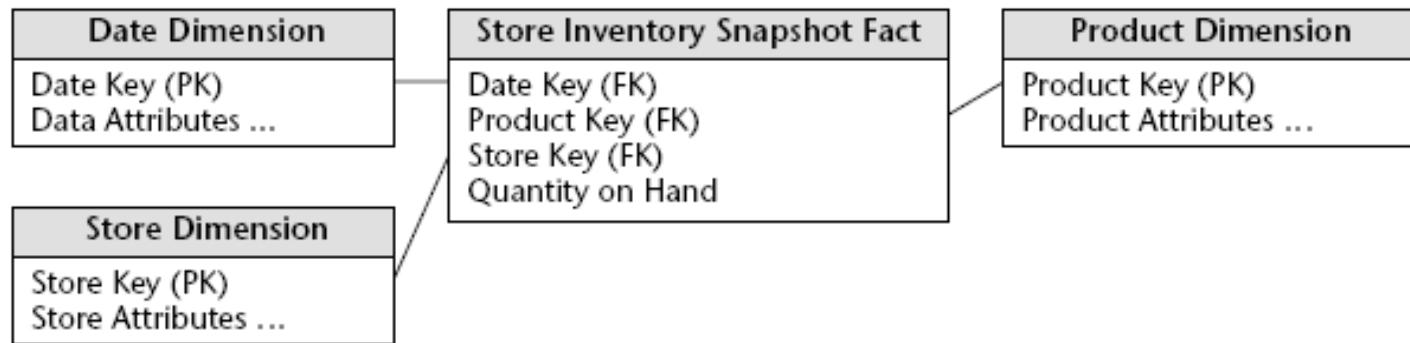
Fatos de transações

- O nível de transação individual representa a visão mais fundamental das operações do negócio. Essas tabelas de fatos representam um evento que ocorreu num ponto instantâneo do tempo.

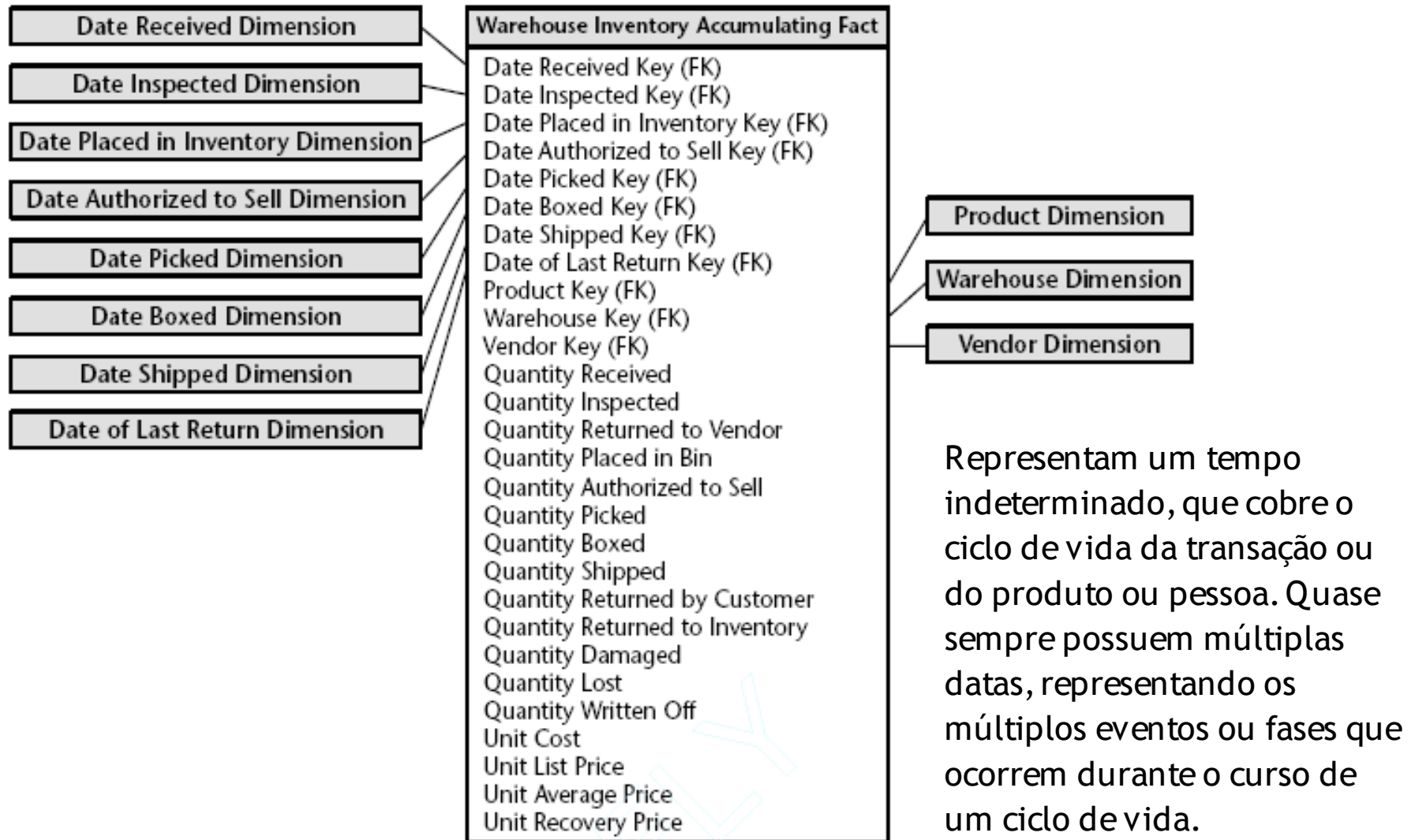


Fatos Instantâneos Periódicos

- São necessários para observar o desempenho cumulativo do negócio em intervalos de tempo regulares e previsíveis. Diferentemente do fato de transação, onde se carrega uma linha para cada ocorrência de evento, com o instantâneo periódico, tira-se uma fotografia da atividade no fim de um dia, uma semana ou um mês, e repetidamente ao fim de cada período.



Fatos instantâneos acumulados



Representam um tempo indeterminado, que cobre o ciclo de vida da transação ou do produto ou pessoa. Quase sempre possuem múltiplas datas, representando os múltiplos eventos ou fases que ocorrem durante o curso de um ciclo de vida.

Tipos clássicos de fatos

Tabela de Comparação dos Tipos de Fatos

CHARACTERISTIC	TRANSACTION GRAIN	PERIODIC SNAPSHOT GRAIN	ACCUMULATING SNAPSHOT GRAIN
Time period represented	Point in time	Regular, predictable intervals	Indeterminate time span, typically short-lived
Grain	One row per transaction event	One row per period	One row per life
Fact table loads	Insert	Insert	Insert and update
Fact row updates	Not revisited	Not revisited	Revisited whenever activity
Date dimension	Transaction date	End-of-period date	Multiple dates for standard milestones
Facts	Transaction activity	Performance for predefined time interval	Performance over finite lifetime

Agregados (1)

- **Materializar (armazenar) ou não?**
 - Vide síndrome da explosão do volume de dados
- **Crítérios para definição de agregados**
 - Passam pela análise dos principais tipos de informação necessárias e pela dificuldade de se obtê-las diretamente das tabelas granulares.
 - Exemplo:

TDLoja (chave-loja, nome-loja, endereço-loja, cidade, estado, *regiao*)

TDProduto (chave-produto, descricao, marca, *categoria*, tipo-embalagem, departamento)

TDDia (chave-dia, data-completa, dia, *mês*, *ano*, período-fiscal, estação)

TFVendas (chave-loja, chave-produto, chave-dia, valor-vendido-real, custo-real, lucro, qtd-vendida)

Hierarquias de dimensões

REGIÃO → LOJA

CATEGORIA → PRODUTO

ANO → MÊS → DIA

Agregados (2)

- Combinações possíveis
 - Ternárias: LOJA X PRODUTO X DIA
→ $2 \times 2 \times 3 = 12$ combinações
 - Binárias:
 - » LOJA X PRODUTO + LOJA X DIA + PRODUTO X DIA
→ $2 \times 2 + 2 \times 3 + 2 \times 3 = 16$ combinações
 - Unárias:
 - » LOJA + PRODUTO + DIA
→ $2 + 2 + 3 = 7$ combinações
 - Total = 35 combinações
- Quais deveriam ser materializadas e armazenadas?
- Qual a distribuição de valores agregados por dimensão?
 - Ex: LOJA
SELECT nome-loja, COUNT(*)
FROM TFVendas, TDLoja
WHERE TFVendas.chave-loja = TDLoja.chave-loja
GROUP BY nome-loja

Agregados (3)

- **Cuidados na definição dos agregados**

- Valores aditivos

- » Nem todas as métricas armazenadas nas tabelas granulares são aditivas em todas as dimensões (fatos semi-aditivos ou não aditivos). Isto significa que os atributos das tabelas fatos de agregados poderão ser diferentes das tabelas fatos granulares.

- Precisão

- » Deve-se definir criteriosamente a precisão dos valores aditivos de agregados, que deverão ser maiores do que os usados nos respectivos valores das tabelas granulares (para evitar overflow na adição)

- » Fatos e dimensões agregados devem estar em tabelas fisicamente diferentes das tabelas granulares, mesmo que o número de tabelas cresça muito. Ferramentas de análise (OLAP, por exemplo) possuem mecanismo de navegação de agregados que escondem a complexidade da estrutura.

Agregados (4)

- **Exemplos**

- Agregação por loja, para todos os produtos, todos os dias.
- Agregação por loja, por mês, para todos os produtos.
- Agregação por região de venda, por mês, por categoria.

Agregados (4)

- **Exemplos**

- Agregação por loja, para todos os produtos, todos os dias.

```
INSERT INTO AG-LOJA AS
```

```
SELECT nome-loja, sum(valor-vendido-real), sum(custo-real)
```

```
FROM TDLoja, TFVendas
```

```
WHERE TDLoja.chave-loja=TFVendas.chave-loja
```

```
GROUP BY nome-loja
```

- Agregação por loja, por mês, para todos os produtos.

```
INSERT INTO AG-LOJA-MÊS AS
```

```
SELECT nome-loja, mês, sum(valor-vendido-real), sum(custo-real)
```

```
FROM TDLoja, TFVendas, TDDia
```

```
WHERE TDLoja.chave-loja=TFVendas.chave-loja AND
```

```
TFVendas.chave-dia=TDDia .chave-dia
```

```
GROUP BY nome-loja, mês
```

Agregados (5)

- **Exemplos**

- Agregação por região de venda, por mês, por categoria.

```
INSERT INTO AG-REG-CAT-MES AS
```

```
SELECT região, mês, categoria, sum(valor-vendido-real), sum(custo-real)
```

```
FROM TDLoja, TFVendas.TDProduto, TDDia
```

```
WHERE TDLoja.chave-loja=TFVendas.chave-loja AND
```

```
TFVendas.chave-dia=TDDia .chave-dia AND
```

```
TFVendas.chave-produto=TDProduto.chave-produto
```

```
GROUP BY região, mês, categoria
```

- **Cuidados operacionais**

- Modelos separados (agregados e granulares) para evitar contenções mútuas no momento de carga ou atualização.
- Carga total versus Atualização incremental: Tempo de processamento versus Complexidade de programas
- Carga/atualização pode requerer processamento paralelo, para otimização

- **Utilização de agregados**

- Navegador de agregados: camada de interface entre a ferramenta OLAP e o servidor de DW. O navegador realiza transparentemente a conversão de comandos SQL granulares nos equivalentes que trabalham informações agregadas.

Dez Erros Comuns a Evitar em Modelagem Dimensional (1)

- **Erro 10:** Colocar atributos de texto usados para restrições e agrupamento numa tabela de fatos.
- **Erro 9:** Limitar atributos descritivos verbosos em dimensões para economizar espaço.
- **Erro 8:** Separar hierarquias e níveis de hierarquia em dimensões múltiplas.
- **Erro 7:** Ignorar a necessidade de cuidar de mudanças em atributos de dimensões.
- **Erro 6:** Resolver todos os problemas de desempenho de consultas adicionando mais hardware.

Dez Erros Comuns a Evitar em Modelagem Dimensional (2)

- **Erro 5:** Usar chaves operacionais ou “inteligentes” para junções de tabelas de dimensão com tabela de fatos.
- **Erro 4:** Negligenciar a declaração e depois a consistência com o grão da tabela de fatos.
- **Erro 3:** Projetar o modelo dimensional baseado em um relatório específico.
- **Erro 2:** Esperar que usuários consultem dados de nível atômico mais baixo num formato normalizado.
- **Erro 1:** Falhar em conformar fatos e dimensões através de diferentes data marts.