

BUSINESS INTELLIGENCE (BI)

MODELAGEM DIMENSIONAL

FERNANDO BARBOSA LIMA



3

LISTA DE FIGURAS

Figura 3.1 – Consulta a dimensão data.....	9
Figura 3.2 – Modelo dimensional sobre vendas	10
Figura 3.3 – Star Schema	13
Figura 3.4 – Star Schema – OLAP Cube	14
Figura 3.5 – Exemplo de Snowflake	15
Figura 3.6 – Técnica 5w3h	19
Figura 3.7 – Dimensões primárias.....	20
Figura 3.8 – Fatos identificados	21
Figura 3.9 – Dimensão data	23
Figura 3.10 – Dimensão produto	24
Figura 3.11 – Dimensão promoção	25
Figura 3.12 – Dimensão loja.....	26
Figura 3.13 – Dimensões vendedor e cliente	26
Figura 3.14 – Fato_vendas.....	27

SUMÁRIO

3 MODELAGEM DIMENSIONAL	4
3.1 A Importância da Modelagem de Dados	4
3.2 Abordagens Técnicas.....	5
3.2.1 Prós e contras do MER em um DW.....	6
3.2.2 Modelagem Dimensional	7
3.2.2.1 Conceitos Básicos	7
3.2.2.2 Exemplo de Modelo Dimensional	10
3.2.2.3 Surrogate Keys, Natural Keys e Smart Keys.....	11
3.2.3 Star Schema.....	12
3.2.4 Star Schema, Cubes, OLAP, ROLAP e MOLAP	13
3.2.5 Snowflake.....	15
3.3 Boas práticas em projetos de DW/BI.....	16
3.4 Processo de Design Dimensional.....	16
3.4.1 Aplicando em um cenário hipotético.....	17
3.4.2 Selecione o processo	17
3.4.3 Determine o grão.....	18
3.4.4 Identifique as dimensões.....	19
3.4.5 Identifique os fatos	20
3.4.6 Atributos comuns para a dimensão data	22
3.4.7 Atributos comuns para a dimensão produto	23
3.4.8 Atributos comuns para a dimensão promoção	24
3.4.9 Atributos comuns para a dimensão loja.....	25
3.4.10 Atributos comuns para a dimensão vendedor e cliente	26
3.4.11 A dimensão degenerada	27
3.4.12 Estrela ou centopeia.....	27
3.5 Dimensões que mudam lentamente (Slowly Changing Dimensions)	28
3.6 Dimensões que mudam rapidamente (<i>Rapidly Changing Monster Dimensions</i>).....	28
3.7 Tipos especiais de dimensões e tabelas.....	29
Conclusão	29
REFERÊNCIAS.....	31
GLOSSÁRIO	33

3 MODELAGEM DIMENSIONAL

3.1 A Importância da Modelagem de Dados

Para entendermos a importância da Modelagem de Dados em um Data Warehouse, é fundamental relembramos duas colocações importantes, detalhadas a seguir.

Inmon, considerado o pai do Data Warehouse, define DW como um conjunto de dados de apoio às decisões gerenciais, integrado, não volátil, variável em relação ao tempo e baseado em assuntos.

- Integrado: os dados são coletados a partir de uma variedade de fontes e fundidos em um todo coerente.
- Não volátil: nenhum dado pode ser alterado ou excluído no DW. Qualquer consulta a um dado relativo a um período de tempo sempre produzirá o mesmo resultado; nenhum dado será excluído enquanto não se tornar obsoleto para o negócio. Em soluções de BI específicas para alguns tipos de negócio, isso pode ocorrer.
- Variável em relação ao tempo: todos os dados no data warehouse são identificados com um período de tempo particular. O DW é focado na manutenção do histórico.
- Orientado ao assunto: possui dados que fornecem informações sobre um assunto específico em vez de fazê-lo sobre as operações em curso de uma empresa.

Segundo Kimball (2011), “Ao processo de preparar os dados de um sistema de informação operacional de forma a se ter uma fonte de informações que possam dar suporte à tomada de decisões deu-se o nome de Data Warehousing”.

Os maiores influenciadores sobre DW, Inmon e Kimball (2011), citam dados e “apoio” ou “suporte” a decisões, nas respectivas definições de Data Warehouse. Juntando as colocações, percebemos que tais dados provêm de sistemas de informações operacionais e devem ser preparados, para serem armazenados por

assuntos, de forma não volátil e variável em relação ao tempo, para servirem de fonte de informações, utilizadas para suporte a decisões.

Portanto, a Modelagem de Dados é, por necessidade, parte fundamental da estrutura e do processamento de uma solução DW, pois o modelo deve ser capaz de comportar o armazenamento de todos os dados relevantes, para apoio à tomada de decisões, bem como, prover alto desempenho de acesso, requisito fundamental para processamentos analíticos complexos.

Neste capítulo, apresentaremos duas técnicas básicas de modelagem para DW, a já conhecida Modelagem Entidade Relacionamento (MER) e a Modelagem Dimensional. Apresentaremos as diferenças entre as duas abordagens e nos aprofundaremos na técnica Dimensional, por ser a abordagem amplamente adotada para modelagem de soluções DW e DM na atualidade.

3.2 Abordagens Técnicas

Existem duas abordagens principais para a modelagem de dados em um DW. Inmon defende a criação de modelos baseados em entidade relacionamento. Kimball (2011), por sua vez, propôs a técnica de modelagem dimensional, basicamente formada por tabelas dimensionais e fato.

No capítulo sobre BI, comentamos que as soluções de Inteligência de Negócios evoluíram rapidamente para o conceito de autoatendimento, possibilitando aos usuários sem conhecimento técnico selecionar dados de diversas fontes e combiná-los de forma livre, para fins de análise, geração de relatórios dinâmicos e tomada de decisões. Também comentamos que, para dar aos usuários tal capacidade, simplicidade é fundamental. Para navegar, selecionar e cruzar os dados sozinhos, os usuários precisam compreender a modelagem do banco de dados, sem grandes dificuldades.

3.2.1 Prós e contras do MER em um DW

O modelo Entidade-Relacionamento é uma técnica de modelagem de Banco de Dados que tem como objetivos o armazenamento estruturado, otimizado, consistente e, por fim, com o menor nível de redundância de dados possível.

Devido às regras de normalização e à criação das entidades atendendo a tais regras, geralmente, a interpretação e o entendimento de modelos ER são realizados com facilidade apenas por especialistas.

Modelagens ER possuem fortíssima semelhança com bancos de dados modelados para o controle de transações diárias. A técnica ER aplica, no DW, os mesmos elementos básicos aplicados nos sistemas OLTP: entidades, relações entre entidades e atributos.

Conceitos como herança e mecanismo de constraint também são empregados. Podemos dizer que apenas uma característica que é minimizada, como uso em modelos ER para sistemas OLTP, é maximizada para o uso em DW, o emprego de atributos calculados. Esse tipo de atributo reduz a necessidade de cálculos em tempo de processamento, aumentando a performance de consultas complexas.

Como já vimos, nesse tipo de banco de dados, o acesso é realizado incluindo atributo a atributo, por meio de queries SQL, geralmente com quantidades elevadas de junções entre tabelas e perda de performance a cada junção adicionada na query.

Não existe uma adequação que facilite a navegação do usuário pelas entidades, requerendo que eles possuam habilidades técnicas. Ponto contrário à tendência crescente no mercado de adotar soluções de autoatendimento para suporte a decisões.

Nem todos os pontos são negativos, modelos ER para DW por natureza controlam a redundância de dados, fato que, se bem explorado, pode gerar uma economia de espaço físico. Outro aspecto comum é que, adotando essa abordagem de modelagem, o reaproveitamento do modelo OLTP é maior, em poucos aspectos, reduzindo os esforços da equipe de modelagem. Perceba que os pontos positivos são essencialmente técnicos e não geram valor ao usuário da solução DW, mais um

fator para o mercado adotar a Modelagem Dimensional como abordagem predominante.

3.2.2 Modelagem Dimensional

Kimball afirma que a Modelagem Dimensional é uma técnica de design diferente do tradicional modelo ER, utilizado nos sistemas de caráter operacional. A abordagem dimensional tem como objetivos criar bancos de dados para suporte a decisão que apresentem os dados de uma maneira padronizada, intuitiva e que permitam acesso de alto desempenho.

A modelagem dimensional é amplamente aceita como abordagem preferida em soluções DW, porque atende dois requisitos importantes: a entrega de dados compreensíveis para os usuários; e a provisão de desempenho em consultas. É uma técnica para elaborar modelos de dados como um conjunto de medidas que são descritas e acessadas por aspectos comuns do negócio. É especialmente útil para organizar, consultar, filtrar, sumarizar, detalhar e suportar a análise de dados.

3.2.2.1 Conceitos Básicos

A modelagem dimensional possui três conceitos básicos, indicados a seguir.

1) Fato:

Fato é uma coleção de dados relacionados que representam um evento do negócio, usado em um DW ou DM para análise e tomada de decisões empresariais. Em um Data Warehouse ou Data Mart, cada fato é registrado em uma linha de uma tabela também chamada de Fato, que trata um processo específico do negócio.

Como exemplo, em um modelo que abrange o processo de Vendas, a tabela Fato pode receber o nome de FATO_VENDAS e deve possuir em cada linha, um item de uma venda, armazenado principalmente em atributos numéricos conhecidos como medidas. Neste exemplo, o item de venda é chamado de grão.

Segundo Kimball (2011), definir o grão é o passo fundamental de um design dimensional. O grão deve estabelecer exatamente o nível de detalhe que será armazenado em uma linha da Tabela Fato. Portanto, o grão deve ser definido antes

de escolhermos as dimensões e deve ser o mais atômico possível (Falaremos sobre grão mais à frente).

2) Medidas:

Uma medida é um atributo numérico de um fato. Por exemplo, em um processo de Vendas, cada item vendido deve ser armazenado em uma linha na tabela Fato e as medidas comuns para esse evento de negócio são: a quantidade vendida, o preço unitário, o custo unitário, o desconto concedido, o valor total, o custo total, o desconto total, o lucro bruto, o valor de imposto, entre outros.

Na maioria dos casos, as medidas, também conhecidas como métricas são aditivas, ou seja, nelas podem ser aplicadas as operações de soma, subtração e média, cruzando-se a seleção por qualquer dimensão (ex.: soma do lucro bruto, de um determinado produto, vendido no ano de 2016).

Existem também as métricas semiaditivas, ou seja, métricas cuja sumarização só faz sentido em alguns casos. Imagine a medida quantidade em estoque de um produto, faz sentido você somar o estoque de hoje, com o estoque de amanhã? A medida estoque reflete a situação do estoque no dia, sendo assim, não faz sentido algum somarmos o estoque de um produto hoje com o de amanhã. Por outro lado, podemos somar a medida estoque de um produto, se selecionarmos o mesmo dia e juntarmos na seleção mais de uma Loja. Concorde?

Por fim, as medidas não aditivas. Percentuais são um bom exemplo desse tipo de medida, imagine que em nossa FATO_VENDAS temos uma métrica de % de lucro bruto, não faz sentido algum somar os percentuais de hoje com os de amanhã.

Existem questionamentos se medidas não aditivas devem ser armazenadas, face ao valor analítico limitado, mas não podemos nos esquecer de que elas geralmente são impressas em relatórios e utilizadas como filtros. Outras dúvidas devem estar pairando em sua cabeça. Todas essas medidas são de qual produto? Vendido para qual cliente? Vendido por qual colaborador? Em qual Loja? Essas questões são respondidas pelas dimensões.

3) Dimensões:

São tabelas que fornecem a base para filtrarmos e analisarmos as medidas da tabela Fato, permitindo a visualização por aspectos diferentes.

As dimensões normalmente respondem às questões do tipo “Quem?”, “O quê?”, “Quando?”, “Onde?” e “Por quê?”. Elas são os parâmetros ou filtros sobre os quais queremos realizar *Online Analytical Processing* (OLAP). Por exemplo, em um modelo dimensional para a análise de vendas, dimensões comuns são: data, produto, cliente, loja, promoção, vendedor.

As tabelas dimensionais armazenam descrições textuais para que o usuário identifique facilmente os registros que ele procura e os utilize como filtros e agrupamentos de análise em ferramentas de BI.

Tabelas dimensionais não são normalizadas e são constituídas por atributos que podem ser arranjados em hierarquias. Uma hierarquia, quando aplicada em uma dimensão, se prevalece da não normalização para facilitar a vida do usuário, que, por natureza, está acostumado com o conceito de agrupamento hierárquico.

Dois exemplos muito comuns são hierarquias de datas e endereços:

- Datas: ano, semestre, trimestre, bimestre, mês, semana e dia da semana.
- Endereço: região, estado, cidade, zona, bairro e endereço.

Para exemplificarmos, o resultado de uma query a partir de um dia aleatório como 03/06/2000, em uma dimensão Data, retornaria algo como mostra a Figura “Consulta a dimensão data”.



	SK_DATA	DATA	DATA_EXTENSO	ANO	SEMESTRE	TRIMESTRE	BIMESTRE	MES	FERIADO	DIA
1	20000603	03/06/00	sábado, 3 de junho de 2000	2000	1	2	3	6 N		3
2	20000604	04/06/00	domingo, 4 de junho de 2000	2000	1	2	3	6 N		4
3	20000605	05/06/00	segunda-feira, 5 de junho de 2000	2000	1	2	3	6 N		5
4	20000606	06/06/00	terça-feira, 6 de junho de 2000	2000	1	2	3	6 N		6
5	20000607	07/06/00	quarta-feira, 7 de junho de 2000	2000	1	2	3	6 N		7

Figura 3.1 – Consulta a dimensão data

Fonte: Elaborado pelo autor (2017)

É muito comum ter cargas de mais de 20 anos em uma dimensão Data, cada 20 anos geram aproximadamente 7.300 linhas na tabela, cada linha será um dia do período inserido. As cargas dessa tabela e das demais do Modelo Dimensional serão vistas no capítulo sobre ETL.

3.2.2.2 Exemplo de Modelo Dimensional

Quando comentamos sobre medidas aditivas, utilizamos como exemplo a obtenção da somatória do lucro bruto, de um determinado produto, vendido no ano de 2016. Essa consulta será facilmente realizada pelo usuário, quando ele selecionar o produto por meio do atributo nm_produto, da DIM_PRODUTO e o ano de 2016, filtrando-o pelo atributo ano, da DIM_DATA. Em uma solução de BI, tal operação será realizada por uma ferramenta, dispensando quaisquer conhecimentos técnicos por parte do usuário. O importante é que o Modelo seja simples e facilmente navegável.

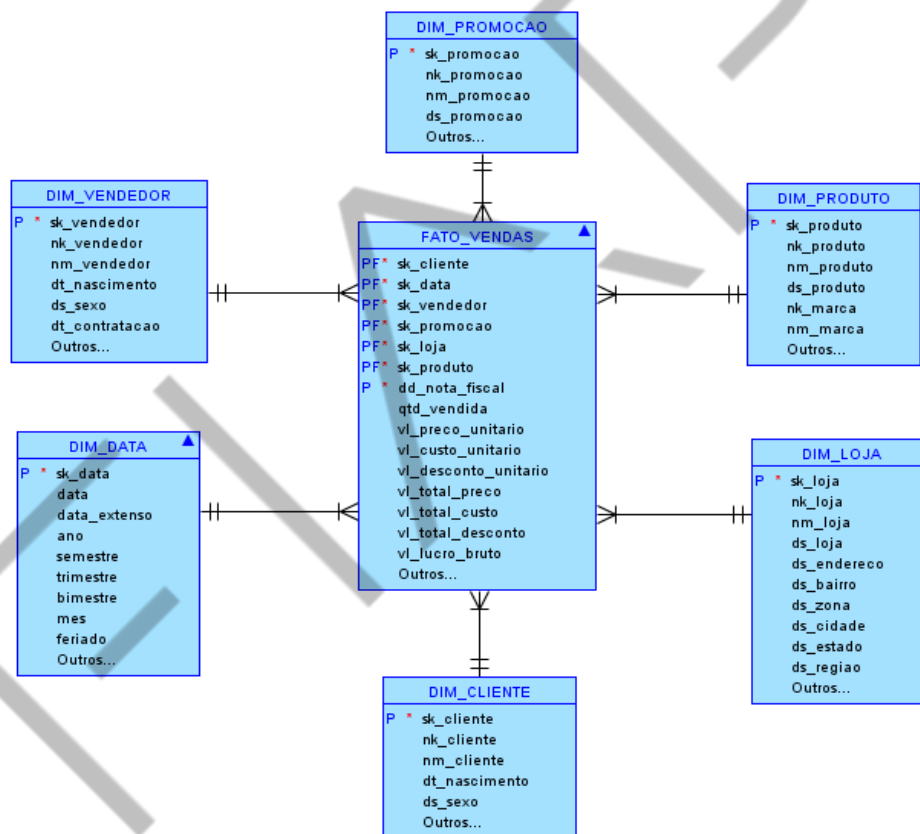


Figura 3.2 – Modelo dimensional sobre vendas
Fonte: Elaborado pelo autor (2017)

Olhando o modelo, é fácil perceber que Modelos Dimensionais, implementados em Banco de Dados Relacionais, possuem as mesmas características da modelagem ER, mas com os três conceitos básicos (Fato, Medidas e Dimensões) empregados como restrições.

Todo modelo dimensional possui no mínimo uma tabela com chave composta, chamada de tabela Fato, relacionada com um conjunto de tabelas chamadas de Dimensões. Cada dimensão deve possuir uma chave primária simples (Surrogate

Key), que corresponde a uma das chaves estrangeiras, que juntas formam a chave composta da tabela Fato.

3.2.2.3 Surrogate Keys, Natural Keys e Smart Keys

1) Surrogate Keys:

Analisando o Modelo Dimensional de exemplo, percebemos que as chaves primárias das dimensões são simples e, por nomenclatura adotada, possuem, no nome do campo, o sufixo sk.

Sk ou Surrogate key é uma chave substituta para a chave primária natural dos dados de origem. Uma sk deve ser um identificador único, numérico do tipo inteiro e sequencial, gerado para cada linha de uma entidade dimensional. Em geral, a primeira linha deve ser preenchida com o valor 1, a segunda com o valor 2 e assim por diante.

Sks não têm valor reconhecido pelo cliente, servem exclusivamente para relacionar as tabelas em um modelo dimensional, portanto, não devem ser visíveis e muito menos manipuláveis pelos usuários. Como sabemos, as tabelas do modelo dimensional sofrerão cargas de dados, obtidos no(s) sistema(s) de origem (OLTP). Durante tais cargas, as Surrogate Keys deverão ser geradas automaticamente, pelo programa responsável ou ferramenta de ETL.

2) Natural Keys:

Note que as chaves naturais não foram descartadas, elas são trazidas durante a carga, em conjunto com os demais campos das dimensões e recebem o sufixo nk. As chaves naturais são importantes, pois possibilitam a rastreabilidade entre origem e destino dos dados, bem como, podem ser usadas como filtros, quando reconhecidas pelo usuário que se acostumou a pesquisá-las no sistema OLTP de origem.

3) Smart Keys:

Voltando a nossa atenção para a consulta feita na dimensão Data, podemos perceber outra característica diferente, note que o valor da chave primária é formado

pela própria data invertida. Esse é um exemplo de chave inteligente, pois ela garante unicidade e ainda aporta um significado real, a própria data a que a linha da tabela corresponde.

A principal ideia não é aplicar um filtro direto da tabela fato, uma vez que a chave será exportada para ela como PF, mas, sim, ajudar no processo de um possível particionamento da Fato, quando ela chegar a um número muito grande de registros e houver perda de valor, na análise dos dados mais antigos.

Dimensões degeneradas:

Com os nossos olhos novamente sobre o modelo, podemos identificar um atributo que faz parte da chave composta da tabela Fato, não é uma FK dimensional e possui um sufixo diferente, dd.

O sufixo corresponde ao conceito de Dimensão Degenerada, que, de forma simplificada, é uma dimensão que não possui atributos descritivos relevantes para análise, apenas uma chave natural, absorvida pela tabela Fato.

Pois bem, no sistema de origem, uma venda é modelada com o conceito de tabela pai “NF” e tabela filha “Item da Nota Fiscal”, cada item da Nota Fiscal será carregado pelo ETL e gerará uma linha na tabela FATO_VENDAS. Esse é o nosso grão. Dúvida – como podemos saber a qual NF o item vendido pertence? Podemos analisar os fatos por produto, data, cliente, vendedor, loja e promoção, mas e por NF?

Para esse tipo de situação que não requer uma dimensão só para um atributo, o conceito de Dimensão Degenerada é aplicável. O número da NF correspondente será gravado com as medidas, em cada linha da Fato, pelo processo ETL.

3.2.3 Star Schema

Star Schema ou modelo estrela são termos comumente utilizados como sinônimos de modelos dimensionais, se repararmos novamente em nosso exemplo, a tabela Fato é o centro da estrela e as Dimensões são as suas pontas. Como

schema é um termo mais técnico, star model ou modelo estrela acabaram sendo os nomes mais adotados pelo mercado.

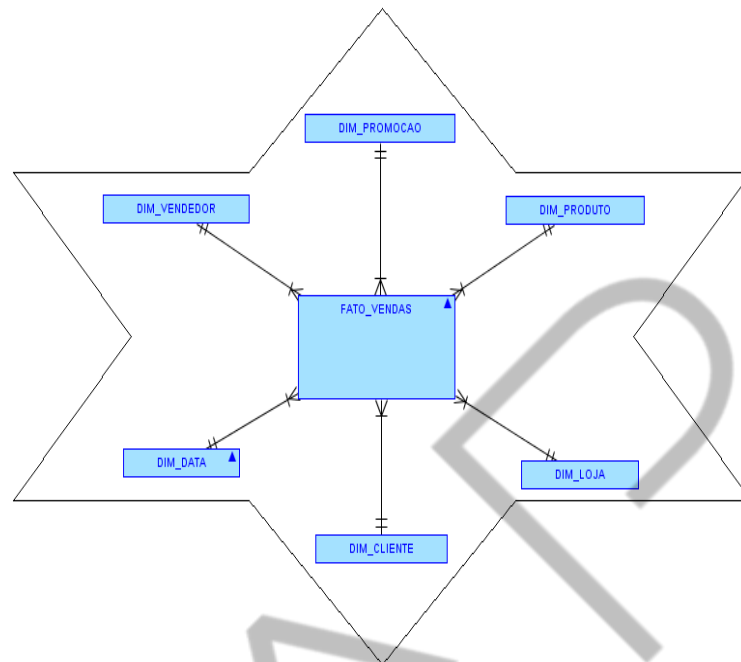


Figura 3.3 – Star Schema
Fonte: Elaborado pelo autor (2017)

3.2.4 Star Schema, Cubes, OLAP, ROLAP e MOLAP

OLAP (Processamento Analítico On-Line) é a capacidade de manipular e analisar um grande volume de dados através de múltiplas perspectivas e, assim, monitorar os fatos e indicadores mais relevantes da organização para a tomada de decisões.

Modelos dimensionais implementados em SGDBs relacionais são conhecidos como modelos estrela, pela semelhança que vimos com o formato de uma estrela. Como o armazenamento dos dados é feito em um banco relacional, esse tipo de implementação leva o nome de ROLAP (Relational On-Line Analytical Processing).

Os modelos dimensionais implementados em banco de dados multidimensionais são conhecidos como cubos MOLAP (Multidimensional On-Line Analytical Processing). Os dados nesse tipo de implementação são armazenados em formatos proprietários de cada ferramenta. As duas implementações OLAP permitem que os próprios usuários explorem, analisem e respondam a perguntas relevantes para o negócio, através de visões dimensionais.

Portanto, os projetos lógicos dimensionais são iguais para os dois tipos de implementação, mas os projetos físicos são diferentes, em face das diferenças de armazenamento de cada ferramenta ou SGBD. Este capítulo foca-se em modelos dimensionais para SGBDs relacionais, mas boa parte do conteúdo pode ser aproveitado em projetos lógicos para bancos multidimensionais.

A abordagem utilizando bancos relacionais possibilita que ferramentas OLAP construam as visões dimensionais a partir dos dados armazenados em modelos dimensionais, utilizando as técnicas de ER, aplicando as restrições que apresentamos até este ponto. Esse tipo de implementação se beneficia da maturidade dos SGBDs em ações importantes, tais como, backup e recovery e escapa de particularidades exclusivas, portanto indesejáveis, de soluções MOLAP.

Soluções MOLAP eram reconhecidas pela maior performance em análises utilizando cubos, mas essa vantagem diminuiu consideravelmente com os avanços em hardware e com o aumento exponencial de dados, mas ainda existem características positivas como capacidades mais ricas de análise e suporte a hierarquias muito complexas, mas tais vantagens variam muito de ferramenta para ferramenta.

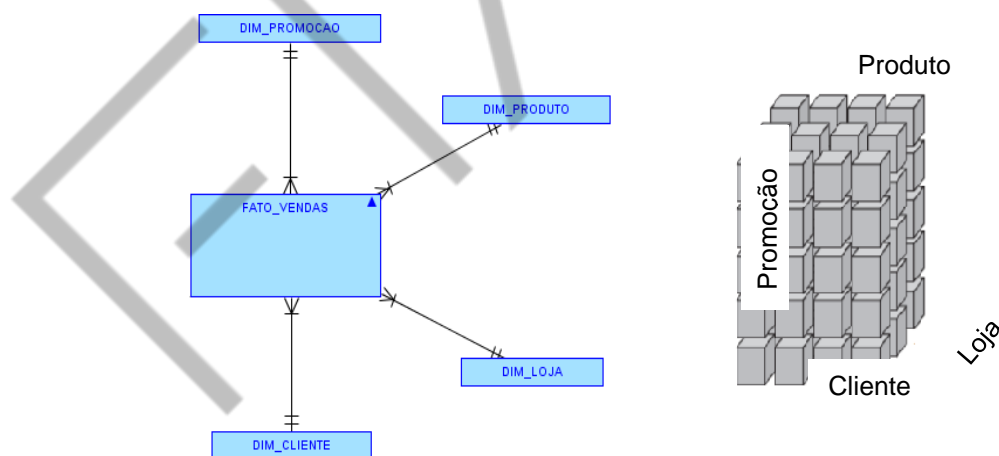


Figura 3.4 – Star Schema – OLAP Cube
Fonte: Elaborado pelo autor (2017)

Comparando graficamente cada abordagem, percebemos que as pontas da estrela são as faces do cubo e, cada vez que o usuário busca uma análise por um atributo de uma dimensão ou por um campo da face do cubo, ele acessa uma perspectiva dimensional sobre os dados que serão analisados.

Como comentamos, independentemente da implementação, soluções OLAP permitem que os próprios usuários explorem, analisem por perspectivas e respondam a perguntas do tipo:

- Qual loja vendeu mais em valores financeiros neste ano?
- Qual loja vendeu mais em valores financeiros, no primeiro trimestre deste ano, por região?

3.2.5 Snowflake

Snowflake ou floco de neve é uma variação do modelo estrela que normaliza uma ou mais dimensões que tenham hierarquia interna, com o objetivo de economizar espaço. Esta variação permite que existam tabelas complementares que se relacionam com as dimensões, além dos relacionamentos existentes entre dimensões e fato. O floco de neve ainda é um modelo dimensional, mas as dimensões seguem a terceira forma normal (3NF) da técnica ER.

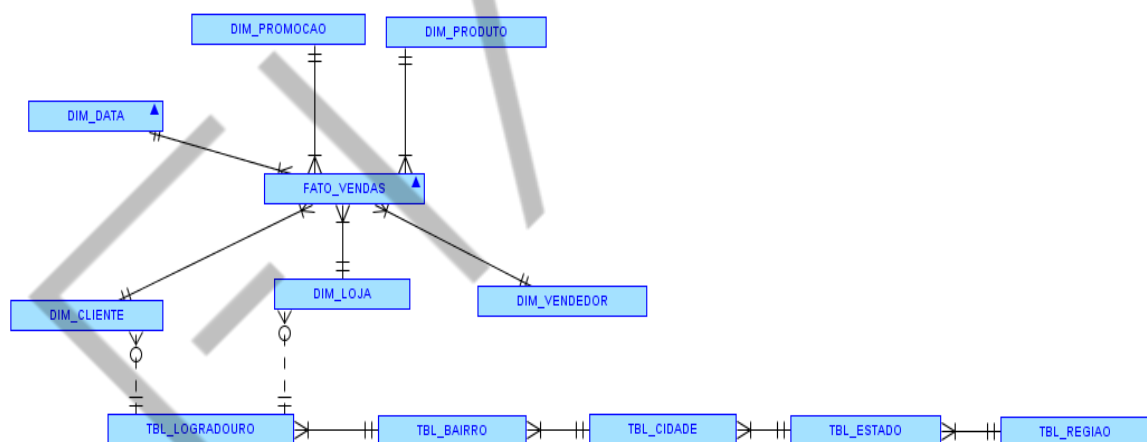


Figura 3.5 – Exemplo de Snowflake
Fonte: Google Imagens (2015)

O modelo floco de neve reduz o espaço de armazenamento dos dados dimensionais, mas acrescenta várias tabelas ao modelo, deixando-o mais complexo para os softwares que utilizarão o banco de dados dimensional e mais complexa a navegação do usuário final pelo modelo. De fato, muitos especialistas argumentam que a economia representada pela redução do espaço de armazenamento não é significativa, pelo tamanho economizado ou pela redução de custos de armazenamento.

Outro fator importante é que mais tabelas serão utilizadas para executar uma consulta, então, mais JOINS SQL serão feitos, impactando a performance. Nos dias de hoje, performance X armazenamento possui forte tendência pela priorização da performance em momentos de tomada de decisão.

3.3 Boas práticas em projetos de DW/BI

No livro *The Data Warehouse Toolkit* (2011), o uso de práticas de desenvolvimento ágeis é encorajado, pois muitos dos princípios fundamentais estão alinhados com as melhores práticas de Kimball, tais como:

- Foco em entregar valor ao negócio. Esse tem sido o mantra de Kimball por décadas.
- Potencializar a colaboração entre a equipe de desenvolvimento e partes interessadas no negócio.
- Comunicação face a face, feedback e constante priorização com as partes interessadas no negócio.
- Adaptar-se rapidamente às inevitáveis evoluções de requisitos.
- Realizar o desenvolvimento de forma iterativa e incremental.

3.4 Processo de Design Dimensional

Em um projeto típico, antes de iniciar a modelagem dimensional, a equipe que realizará o trabalho precisa questionar e compreender muito bem as necessidades do negócio, bem como identificar as reais condições dos dados de origem.

Os requisitos de negócio devem ser levantados por meio de reuniões de entendimento, envolvendo tomadores de decisão ou seus representantes, para que a equipe de modelagem entenda os objetivos de análise, utilizando como base indicadores de desempenho. A equipe deve levantar os processos de negócio envolvidos, compreender a dinâmica das tomadas de decisão, de acordo com o nível gerencial ou natureza do objeto de análise e identificar as necessidades

analíticas de suporte, para que as decisões sejam tomadas da melhor forma possível.

Em paralelo, a cada objetivo de análise entendido, consultas devem ser feitas às documentações e aos especialistas dos sistemas OLTP selecionados como fonte dos dados de origem, para avaliar a existência e as reais possibilidades de extração desses dados.

Ainda se apoiando em Kimball (2011), o autor sugere um conjunto de técnicas que orienta em quatro passos as principais decisões que devem ser feitas durante o processo de design de uma modelagem dimensional.

Os quatro passos sugeridos são:

- Selecione o processo.
- Determine o grão.
- Identifique as dimensões.
- Identifique os fatos.

3.4.1 Aplicando em um cenário hipotético

Imagine uma rede com 50 lojas de departamentos espalhadas em quatro regiões do país, cada loja possui aproximadamente 30.000 produtos disponíveis para venda; em decorrência da crise, a gerência está preocupada em aumentar as vendas e maximizar o lucro. A rede de lojas se utiliza de promoções que ofertam descontos, publicadas e divulgadas em panfletos, nos sistemas de som das lojas e nas rádios locais.

Saber o que está dando certo é muito importante e, nesse momento, deve ser potencializado.

3.4.2 Selecione o processo

Ao primeiro passo, cabe a decisão de qual ou quais processos é possível modelar, por meio do cruzamento dos requisitos de negócios para análise, com o levantamento dos dados disponíveis.

O primeiro modelo deve focar as questões mais críticas (FCS) para os usuários, desde que factíveis. Ser factível envolve diversas considerações, tais como: disponibilidade e qualidade dos dados e o compromisso da organização em relação ao contexto.

No cenário descrito, a questão mais crítica identificada com os tomadores de decisão foi um melhor entendimento das compras dos clientes. Sendo assim, o processo de negócio que será modelado é a transação de vendas, possibilitando a análise de quais produtos estão vendendo mais, em quais lojas, em que períodos e em quais condições promocionais.

3.4.3 Determine o grão

Após a identificação do processo, encontramos uma importante e crítica tomada de decisão sobre granularidade dos dados.

Qual o nível de detalhe sobre vendas que ficará disponível no modelo dimensional?

Temos de manter em mente o objetivo de prover aos usuários o nível de detalhe mais atômico, possível de ser capturado no processo de negócio que selecionamos para modelar.

Segundo Kimbal (2011), quanto mais detalhada e atômica for a medida dos fatos, mais coisas saberemos com certeza. No processo escolhido, o grão adequado é o item vendido, ou seja, o item da nota fiscal, pois nele está o maior nível de detalhe da venda.

Se, por engano, escolhêssemos como grão a nota fiscal e não o item da nota, não conseguiríamos descobrir os produtos que mais venderam, pois eles são relativos ao item da nota e não à nota fiscal. Lembrando que uma nota fiscal é composta por um ou mais itens e cada item possui um produto, vendido em uma quantidade igual ou superior a 01.

Apesar do que explicamos, em outras situações, a equipe pode, por conta e risco, escolher uma granularidade mais alta, uma agregação de um dado atômico. Entretanto, a escolha do grão em um nível mais alto limitará a modelagem a

relacionar um número menor de dimensões e/ou envolver dimensões menos detalhadas.

Uma escolha desse tipo, com certeza, deixará o modelo dimensional vulnerável e insuficiente, caso os usuários apresentem novos requisitos, com maior profundidade nos detalhes em que eles pretendem mergulhar.

Nesses casos, Kimbal (2011), salienta que os usuários se defrontarão com um muro analítico, pois será impossível visualizar os detalhes dos dados nas consultas, uma vez que serão extraídos dos sistemas OLTP, sumarizados e gravados sem os detalhes no modelo dimensional, pelo processo de ETL. Sendo assim, em nosso exemplo, o grão adequado continua sendo o item da nota fiscal.

3.4.4 Identifique as dimensões

Definido o grão da tabela Fato, a identificação das dimensões primárias é o próximo passo, e elas surgem naturalmente, analisando as perspectivas pelas quais o negócio pretende analisar, os relacionamentos que o grão possui na fonte de origem dos dados e aplicando uma técnica chamada 5w3h, adaptada da área de administração, que funcionará como um mecanismo de melhor entendimento do modelo e de identificação das principais dimensões.

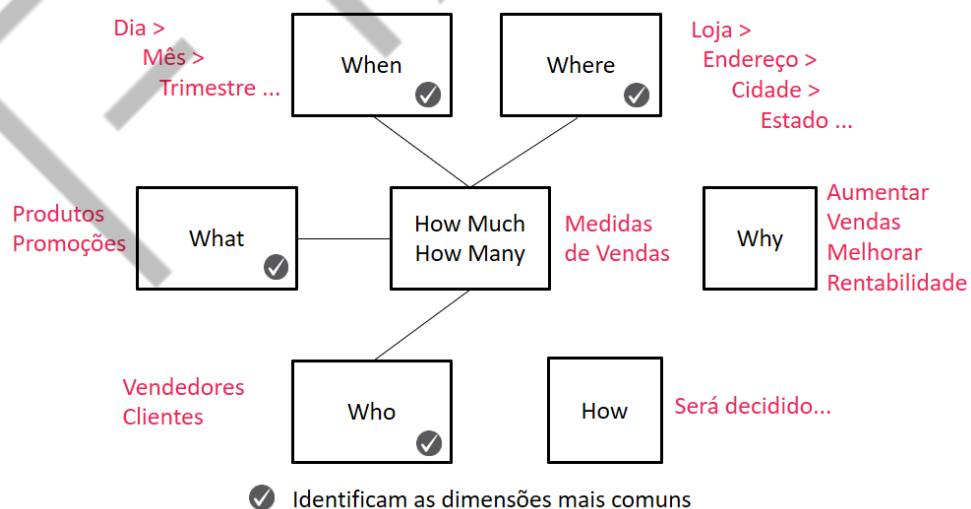


Figura 3.6 – Técnica 5w3h
Fonte: Elaborado pelo autor (2017)

Técnica 5w3h aplicada, encontramos:

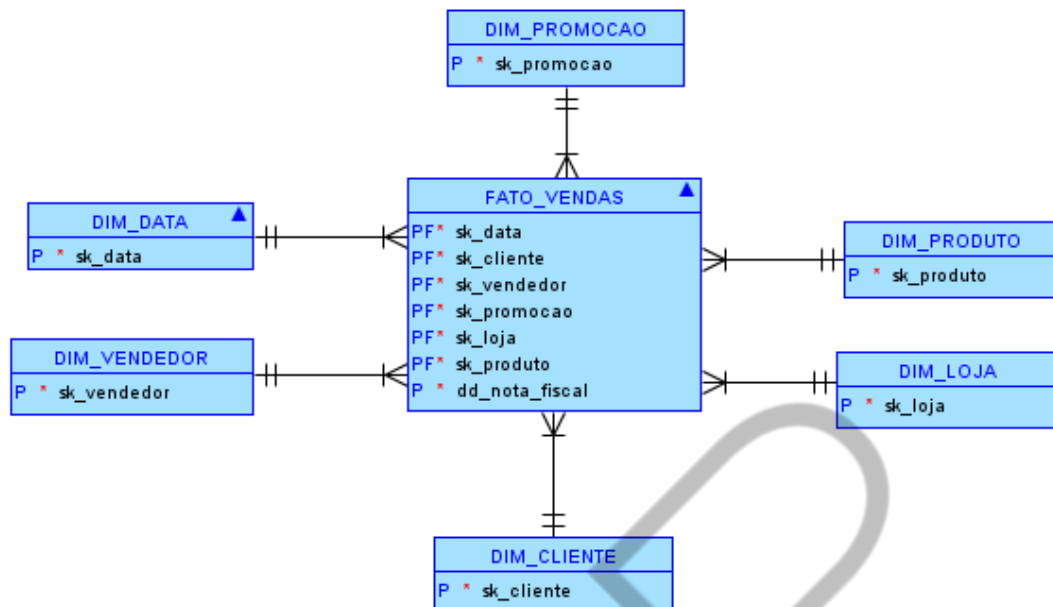


Figura 3.7 – Dimensões primárias
Fonte: Elaborado pelo autor (2017)

Identificadas as dimensões primárias, a sugestão é seguir para o próximo passo, ainda não é o momento de mergulhar na definição de todos os atributos das dimensões primárias. Novas dimensões podem surgir e pior, a equipe pode perder a visão geral do design, se envolvendo nos detalhes. A equipe deve se aprofundar no final do último passo.

3.4.5 Identifique os fatos

O último passo do design é responsável por determinar cuidadosamente os fatos que aparecerão na tabela Fato. Novamente, a declaração do grão ajudará na definição dos atributos que aparecerão na tabela.

Os tomadores de decisão estão especialmente interessados em analisar as métricas de desempenho e os detalhes sobre o que o processo de negócio deve medir deverá ser informado por eles.

Todos os fatos candidatos devem ser verdadeiros para o grão já definido. Fatos que não pertençam ao grão devem ser pensados em uma tabela Fato separada ou gerar uma revisão no grão escolhido. A equipe não deve tentar criar adaptações, elas não funcionarão. Fatos típicos são medidas numéricas, como a quantidade vendida, o preço unitário, o custo unitário, o desconto unitário, o valor total, o custo total, o desconto total e o lucro bruto.

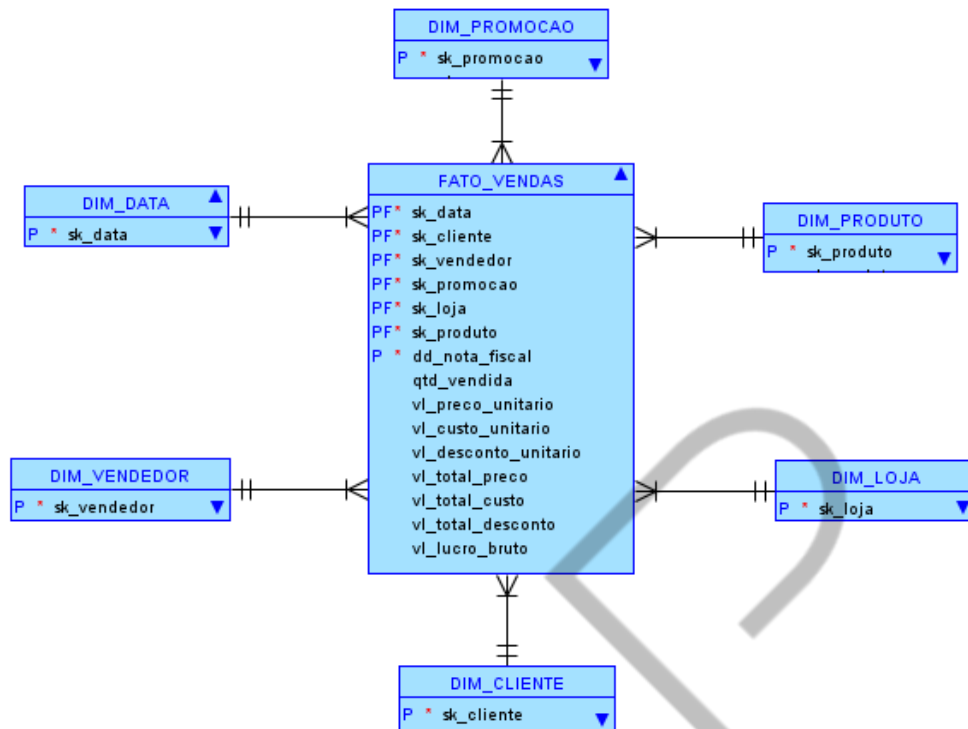


Figura 3.8 – Fatos identificados
Fonte: Elaborado pelo autor (2017)

Os totais serão muito utilizados por análises feitas por todas as dimensões. Os usuários poderão realizar operações de slice e dice pelos atributos das dimensões sem se preocupar, pois cada soma dessas medidas será válida e correta.

Uma recomendação importante é resistir à tentação de modelar dando mais atenção às documentações dos sistemas ou aos especialistas envolvidos com os dados da fonte, do que conversando com as pessoas de negócios. A origem é importante, mas ela não reflete o que o negócio precisa saber.

Por fim, dados calculados devem ser persistidos ou não? Kimball geralmente recomenda que sejam armazenados fisicamente e computados consistentemente pelo ETL, eliminando a possibilidade de que o usuário ou alguém de TI calcule errado. O custo de um usuário que tome decisões sobre um dado calculado erroneamente supera o custo de armazenamento. O armazenamento garante que todos os usuários e aplicações de BI terão referências aos dados calculados de forma consistente.

Ao terminar a identificação dos Fatos, a equipe deve se debruçar sobre os detalhes das dimensões primárias e os detalhes de quaisquer outras dimensões que tenham surgido durante os quatro passos.

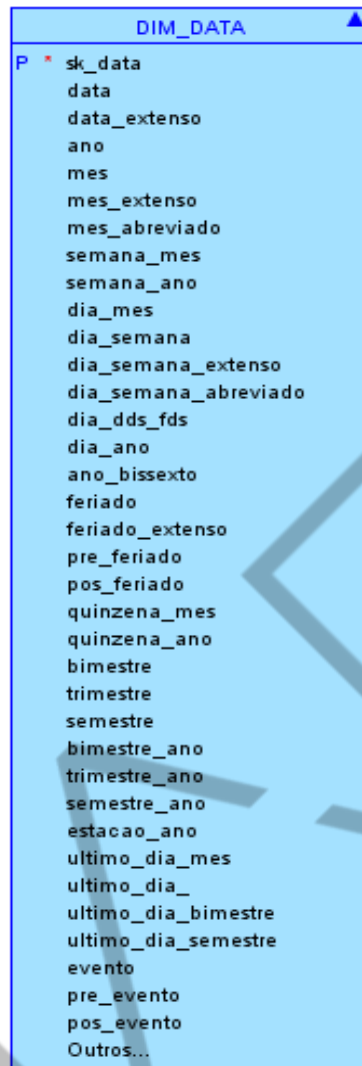
Como estamos trabalhando com um cenário hipotético, vamos apresentar sugestões de atributos mais comuns para as dimensões.

3.4.6 Atributos comuns para a dimensão data

Para que o usuário analise adequadamente todas as medidas, elas devem ser apresentadas em contextos apropriados. Tais contextos sempre possuem um elemento envolvendo datas. Portanto, recomenda-se a criação de uma dimensão para tanto. Alguns autores utilizam como nome dimensão tempo, outros, dimensão data e outros ambos os nomes, dependendo da granularidade necessária para a análise.

Imagine que o contexto analisado precise ser filtrado por dia, mas, em certos casos, por hora do dia. Uma carga de 20 anos, dia a dia, já é considerável. Imagine com 24 linhas para cada dia, se considerarmos tudo na mesma dimensão. Em situações específicas como essa, podemos considerar a criação de duas dimensões: uma representando os dias, no período de décadas; e a segunda, com 24 linhas, representando a quantidade de horas de um dia. Com ambas relacionadas a Fato, podemos saber facilmente o dia e a hora em que o evento ocorreu.

Vários atributos são relevantes de acordo com o negócio, por exemplo, marcar se no dia possui ou faz parte de um evento ou se o dia é anterior ou posterior a um evento podem ser informações muito relevantes. Você acha que a Oktoberfest não influencia a venda de cervejas em Blumenau?



DIM_DATA	
P	sk_data
	data
	data_extenso
	ano
	mes
	mes_extenso
	mes_abreviado
	semana_mes
	semana_ano
	dia_mes
	dia_semana
	dia_semana_extenso
	dia_semana_abreviado
	dia_dds_fds
	dia_ano
	ano_bissexto
	feriado
	feriado_extenso
	pre_feriado
	pos_feriado
	quinzena_mes
	quinzena_ano
	bimestre
	trimestre
	semestre
	bimestre_ano
	trimestre_ano
	semestre_ano
	estacao_ano
	ultimo_dia_mes
	ultimo_dia_
	ultimo_dia_bimestre
	ultimo_dia_semestre
	evento
	pre_evento
	pos_evento
	Outros...

Figura 3.9 – Dimensão data
Fonte: Elaborado pelo autor (2017)

3.4.7 Atributos comuns para a dimensão produto

A DIM_PRODUTO costuma ter muitos registros a mais do que a quantidade de produtos atualmente vendidos, pois contém produtos não mais disponíveis como histórico de vendas.

A hierarquia dos produtos geralmente possibilita vários grupamentos, por exemplo, um produto pode ser agrupado por marca que, por sua vez, pode ser agrupada em categorias, e essas, depois, em departamentos. Esse tipo de análise é bastante usual e ajuda muito os tomadores de decisão, ampliando o entendimento e reduzindo tempo.

Outros atributos que não fazem parte da hierarquia de produtos, geralmente, devem estar presentes, pois são importantes, veja o tipo de embalagem.

Esse atributo pode conter valores, tais como: garrafa, caixa, saco ou outro qualquer, que em determinados contextos podem ser relevantes aos analistas de negócio. Portanto, devemos resistir à tentação de normalizar essa dimensão.

DIM_PRODUTO	
P *	sk_produto
	nk_produto
	nm_produto
	ds_produto
	nk_marca
	nm_marca
	nk_tipo_marca
	ds_tipo_marca
	nk_subcategoria
	nm_subcategoria
	nk_categoria
	nm_categoria
	nr_departamento
	ds_departamento
	ds_tipo_embalagem
	nr_largura_embalagem
	nr_altura_embalagem
	nr_peso
	nr_prazo_validade
	dt_inicio_venda
	dt_fim_venda
	Outros...

Figura 3.10 – Dimensão produto
Fonte: Elaborado pelo autor (2017)

3.4.8 Atributos comuns para a dimensão promoção

Essa é uma das dimensões mais importantes para o nosso cenário hipotético. Relembramos que é por meio dessa dimensão que será possível entender sob quais condições um produto foi vendido.

Podemos realizar, por ela, as seguintes análises:

- Análise do ganho de vendas.
- Quedas de vendas em período anterior ou posterior à promoção.
- Análise de ganho e perda relacionados.
- Análise de ganho total de vendas.
- Análise da lucratividade da promoção.

DIM_PROMOCAO	
P *	sk_promocao
	nk_promocao
	nm_promocao
	ds_promocao
	ds_publico_alvo
	ds_objetivos_promocao
	ds_tipo_promocao
	ds_cupom_desconto
	ds_categoria_promocao
	ds_tipo_midia
	nr_percentual_desconto
	qtd_minima
	qtd_maxima
	dt_inicio
	dt_fim
	vl_custo_promocao
	Outros...

Figura 3.11 – Dimensão promoção
 Fonte: Elaborado pelo autor (2017)

Somente a dimensão Promoção pode nos fornecer maneiras tão ricas de cruzar dados tão diferentes, tais como: “Qual o efeito da primeira promoção que combinou uma redução do preço de vendas, com cartazes na entrada da loja?”.

Perceba, existem várias condições causais possíveis, altamente correlacionadas. Importante, como o relacionamento entre a dimensão promoção e a tabela fato é identificado, devemos criar um registro do tipo “Nenhuma Promoção em Vigor”, para todos os casos de vendas sem nenhuma promoção associada.

3.4.9 Atributos comuns para a dimensão loja

A dimensão Loja armazena todas as lojas da rede que estamos analisando e possibilita a análise por outra perspectiva muito importante, a perspectiva geográfica. Por exemplo, os usuários podem iniciar a análise por uma cidade e fazer o Roll Up para estado, região e país; ou começar pelo país e fazer Drill Down até chegar à cidade.

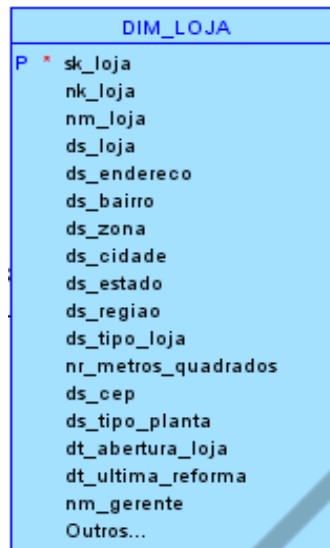


Figura 3.12 – Dimensão loja
Fonte: Elaborado pelo autor (2017)

3.4.10 Atributos comuns para a dimensão vendedor e cliente

Essas dimensões possuem informações sobre os vendedores e os clientes da rede, ambos os casos variam muito em relação à forma como a rede de lojas hipotética trata informações pessoais. Em geral, a dimensão vendedor será utilizada para análise de performance por classe, grau de instrução, departamento, entre outras. Em relação aos clientes, a rede precisa criar mecanismos para conhecê-los, como cartão fidelidade e descontos por perfil. Da mesma forma sugerida na dimensão promoção, devemos criar um registro do tipo “Cliente não identificado”, para todos os casos de vendas nos quais o cliente não queira se identificar.

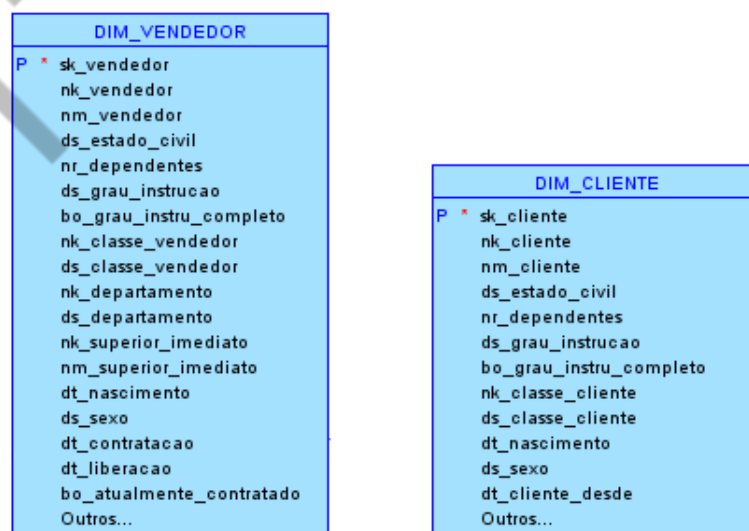
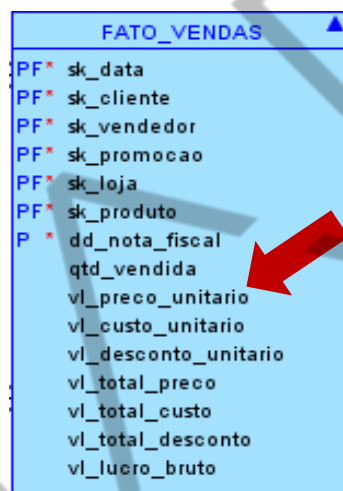


Figura 3.13 – Dimensões vendedor e cliente
Fonte: Elaborado pelo autor (2017)

3.4.11 A dimensão degenerada

O número da nota fiscal, apesar de ser uma perspectiva de análise, não dá origem a uma dimensão. Em casos em que o grão representa uma única transação de um conjunto reconhecido pelo negócio é muito comum ter uma dimensão degenerada, representando tal conjunto.

Números de Nota Fiscal, Número de Pedido e Número da Prescrição Médica, usualmente, dão origem a dimensões vazias que são representadas por meio de dimensões degeneradas em tabelas Fato. No nosso exemplo, temos como DD o número da nota fiscal.



FATO_VENDAS	
PF *	sk_data
PF *	sk_cliente
PF *	sk_vendedor
PF *	sk_promocao
PF *	sk_loja
PF *	sk_produto
P *	dd_nota_fiscal
P	qtd_vendida
P	vl_preco_unitario
P	vl_custo_unitario
P	vl_desconto_unitario
P	vl_total_preco
P	vl_total_custo
P	vl_total_desconto
P	vl_lucro_bruto

Figura 3.14 – Fato_vendas
Fonte: Elaborado pelo autor (2017)

3.4.12 Estrela ou centopeia

Quando identificamos um número muito grande de dimensões, devemos avaliar se entre elas existem dimensões dependentes ou parcialmente independentes. Nessas situações, Kimball recomenda a combinação de duas ou mais dimensões em uma única dimensão e considera como um erro de modelagem a representação de elementos de uma hierarquia como dimensões separadas na tabela de Fatos.

Como parâmetro, qualquer modelo que exceda a 15 dimensões deve ser alvo de melhor análise.

3.5 Dimensões que mudam lentamente (*Slowly Changing Dimensions*)

Imagine uma situação na qual um fabricante altera o tipo de embalagem de um produto. Apesar de não acontecer com frequência, essa mudança impacta a dimensão. O que devemos fazer?

Vamos olhar as três opções mais adotadas para tratar *Slowly Changing Dimensions* ou SCD.

- Atualizar o valor, gravando o novo por cima do valor antigo.

Solução simples, mas que não preserva histórico. Ou seja, se vendíamos leite em saquinho e ele passou a ser engarrafado, ao gravarmos por cima, todas as vendas anteriores serão enxergadas como um produto engarrafado.

- Adicionar uma nova linha com o novo valor do atributo atualizado, mantendo os valores anteriores. Geralmente, adicionando uma coluna para o controle de versão ou duas colunas com datas para controle de período, ou uma com data e uma segunda como flag, sinalizando ativo ou não.

De todas, essa é a técnica mais utilizada para resolver os casos de dimensões que mudam lentamente.

- Adicionar uma nova coluna, preservando o valor anterior e inserindo o novo valor na nova coluna.

Permite a manutenção de duas ou mais visões simultâneas do histórico, mas gera campos nulos e possui um número limitado de possibilidades. Incluir, a toda hora, uma nova coluna, com certeza, é um problema.

3.6 Dimensões que mudam rapidamente (*Rapidly Changing Monster Dimensions*)

Imagine outra situação, desta vez, estamos fazendo um modelo dimensional para uma empresa que pratica mensalmente o rodízio de funcionários por área demográfica. Já conhecemos técnicas para tratar mudanças em dimensões, mas as técnicas que vimos resolvem bem mudanças esporádicas. Pensando na solução

mais adotada, criar uma linha todos os meses para cada funcionário, com o objetivo de registrar a área atual não é uma boa solução. Concorda?

Para esses casos, particionar a dimensão é uma boa solução, deixe os dados estáticos em uma parte e acomode os dados voláteis em uma segunda. Ou seja, crie uma dimensão para tratar apenas os dados demográficos. A aplicação dessa técnica gera dimensões menores, conhecidas como mini dimensões.

3.7 Tipos especiais de dimensões e tabelas

- **Junk Dimension:** um tipo de dimensão abstrata para acomodar a decodificação de flags e indicadores de baixa cardinalidade, evitando que eles fiquem na tabela Fato.
- **Outrigger:** é o nome que se dá a uma tabela ligada a uma dimensão quando se adota o modelo snowflake.
- **Role Playing Dimensions:** uma única dimensão física pode ser referenciada várias vezes em uma tabela de fatos, com cada referência ligando a uma função logicamente distinta para a dimensão. Um exemplo bem comum, uma tabela de fatos pode ser analisada através de várias datas, cada uma será representada por uma chave estrangeira, mas de origem da mesma dimensão data.
- **Bridge Table (tabela ponte):** em um esquema dimensional clássico, cada dimensão anexada a uma tabela de fatos tem um único valor consistente com o grão da Fato. Existem situações em que uma dimensão é legitimamente multivalorada. Por exemplo, um paciente que passa por um exame pode receber vários diagnósticos simultâneos. Em casos como esse, a dimensão multivalorada deve ser anexada à tabela de Fatos, por meio de uma tabela ponte.

Conclusão

Neste capítulo, objetivou-se abordar o básico e essencial sobre Modelagem Dimensional. Apresentamos a importância da modelagem de dados em um DW, as

diferentes abordagens técnicas de modelagem, os conceitos básicos de modelagem dimensional e as boas práticas de projeto.

Tais conteúdos geraram a base necessária para a aplicação das técnicas em um cenário hipotético, por meio do qual foi possível apresentar uma sugestão de processo para design dimensional e exemplificar tabelas fato, dimensões e atributos mais comuns.

Ao final, foram apresentadas técnicas para lidar com mudanças em dimensões e alguns tipos de dimensões especiais. Apesar do conteúdo abrangente, existem diversos assuntos sobre modelagem dimensional que podem ser explorados por você.

Sendo assim, recomendamos a leitura dos inúmeros conteúdos que são abordados nos livros e URLs de referência deste capítulo.

REFERÊNCIAS

BALLARD, Chuck et al. **Dimensional Modeling: In a Business Intelligence Environment**. IBM-International Technical Support Organization, 2006.

_____. **Data modeling techniques for data warehousing**. IBM Corporation International Technical Support Organization, 1998.

GONÇALVES, Marcio. **Extração de dados para Data Warehouse**. Rio de Janeiro: Axcell Books, 2003.

HAHN, Seungrahn et al. Capacity planning for business intelligence applications: **Approaches and methodologies**. IBM Redbooks, 2000.

INMON, William H. **Building the data warehouse**. John Wiley & Sons, 2005.

KIMBALL, Ralph; CASERTA, Joe. **The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data**. John Wiley & Sons, 2011.

KIMBALL, Ralph; ROSS, Margy. **The data warehouse toolkit: the complete guide to dimensional modeling**. John Wiley & Sons, 2011.

KIMBALL GROUP. **Slowly Changing Dimensions**. 2008. Disponível em: <<http://www.kimballgroup.com/2008/08/slowly-changing-dimensions/>>. Acesso em: 15 dez. 2017.

_____. **Design Tip #105 Snowflakes, Outriggers, and Bridges**. 2008. Disponível em: <<http://www.kimballgroup.com/2008/09/design-tip-105-snowflakes-outriggers-and-bridges/>>. Acesso em: 15 dez. 2017.

_____. **Design Tip #113 Creating, Using, and Maintaining Junk Dimensions**. 2009. Disponível em: <<http://www.kimballgroup.com/2009/06/design-tip-113-creating-using-and-maintaining-junk-dimensions/>>. Acesso em: 15 dez. 2017.

_____. **Add Mini-Dimension**. Disponível em: <<http://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/type-4-mini-dimension/>>. Acesso em: 15 dez. 2017.

_____. **Role-Playing Dimensions**. Disponível em: <<http://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/role-playing-dimension/>>. Acesso em: 15 dez. 2017.

_____. **Multivalued Dimensions and Bridge Tables**. Disponível em: <<http://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/multivalued-dimension-bridge-table/>>. Acesso em: 15 dez. 2017.

MACHADO, Felipe Nery Rodrigues. **Tecnologia e Projeto de Data Warehouse:** uma visão multidimensional. Ética, 2006.

REINSCHMIDT, Joerg; FRANCOISE, Allison. Business intelligence certification guide. IBM **International Technical Support Organisation**, 2000.

SINGH, Harry S. **Data warehouse:** conceitos, tecnologias, implementação e gerenciamento. Tradução Mônica Rosemberg. São Paulo: Makron Books, 2001.

TURBAN, Efraim et al. **Business Intelligence:** um enfoque gerencial para a inteligência do negócio. Bookman, 2009.

EMEND

GLOSSÁRIO

DBMS	Database Management System ou Sistema Gerenciador de Banco de Dados. É software que interage com usuários finais, outras aplicações e o próprio banco de dados para capturar e analisar dados.
IBM	International Business Machines (IBM) – é uma empresa dos Estados Unidos voltada para a área de informática. Uma das poucas na área de tecnologia da informação (TI) com uma história contínua que remonta ao século XIX.
OLTP	Processamento de transações on-line (OLTP) descreve a forma como os dados são processados por um sistema informatizado. Sistemas OLTP armazenam seus dados de forma normalizada e, geralmente, processam enormes quantidades de operações CRUD, realizadas pelo usuário final.
DW	Data Warehouse é um conjunto de dados de apoio às decisões gerenciais, integrado, não volátil, variável em relação ao tempo e baseado em assuntos.
DM	Data Mart é um subconjunto de dados corporativos, geralmente focados em assuntos especiais e de valor para um departamento da corporação, unidade corporativa ou conjunto de usuários. Um DM é definido pelo escopo funcional que atende e não pelo seu tamanho. Geralmente, é considerado como subconjunto de um DW.
OLAP	É a capacidade de manipular e analisar um grande volume de dados através de múltiplas perspectivas e, assim, monitorar os fatos e indicadores mais relevantes da organização, por meio de painéis de controle e relatórios executivos desenvolvidos para facilitar a visualização, o entendimento dos fatos e a tomada de decisões.

ETL	Extract Transform Load (Extração Transformação Carga) é o processo de extração, transformação e carga dos dados, oriundos de fontes diversas em modelos dimensionais no DW, para que os usuários finais possam realizar consultas e tomar decisões.
SQL	Structured Query Language é um idioma padrão para armazenar, manipular e recuperar dados em bancos de dados.
Drill down	Ato de fazer uma exploração em diferentes níveis de detalhe ou hierarquias de informações de uma dimensão, partindo da menos detalhada, para a mais detalhada.
Roll up	É o ato inverso ao Drill down, ou seja, uma exploração em diferentes níveis de detalhe ou hierarquias de informações de uma dimensão, partindo da mais detalhada, para a menos detalhada.
Slice	Slice é um filtro que permite ver os dados de diferentes visões, Slice apresenta dados de uma única dimensão de um cubo.
Dice	Dice é um filtro que permite ver os dados de diferentes visões. Apresenta um subcubo ou intersecção de vários slices.