

EDW e Modelagem Dimensional

- Modelagem Dimensional
- Arquitetura de Barramento do Enterprise Data Warehouse
- Matriz de Barramento
- Tabela Fato
- Tabela de Dimensão
- Star Schema
- Tabela Fato sem Fato
- Tabela Fato de Snapshot Periódico
- Fatos Conformados
- Chaves Dimensionais
- Tabela Fato Transacional
- Tabela Fato de Snapshot Acumulado
- Natural Key
- Dimensões de Etapa (Step Dimensions)
- Tabelas Fato Agregadas
- Fatos Aditivos
- Dimensões Multivaloradas e Tabelas Ponte (bridge table)
- Tabelas Fato Consolidadas
- Fatos Agregados e Atributos de Dimensões
- Dimensões Genéricas Abstratas
- Dimensões de comentário
- Dimensões de Auditoria
- Dimensão Conformada
- Dimensões reduzidas (Shrunken Dimensions)
- Slowly Changing Dimensions

- Dimensões Degeneradas
- Mantendo a Granularidade na Modelagem Dimensional
- Pense Dimensionalmente
- Schemas de Eventos de Erros
- Surrogate Key
- Data Profiling

Modelagem Dimensional

Modelagem dimensional é uma disciplina de design que abrange a modelagem relacional formal e a engenharia de realidades textuais e numéricas. Comparada à terceira forma normal da Modelagem Entidade-Relacionamento, é menos rigorosa (permitindo ao designer mais discricão na organização das tabelas), porém mais prática pois acomoda a complexidade de um banco de dados enquanto contribui para a melhoria do seu desempenho. A modelagem dimensional possui um portfólio extenso de técnicas para lidar com situações do mundo real (KIMBALL e ROSS, 2010).

Devido a sua simplicidade na apresentação dos dados, a modelagem dimensional tem sido amplamente aceita como a técnica dominante de modelagem de *Data Warehouses*. A simplicidade é a chave fundamental que permite aos usuários finais navegarem com eficiência sobre os dados apresentados pelo DW. Ao focar consistentemente em uma perspectiva orientada para os negócios, recusando-se a comprometer a objetivos específicos de usuários, você estabelece um design coerente que atende às necessidades analíticas da organização (KIMBALL e ROSS, 2013).

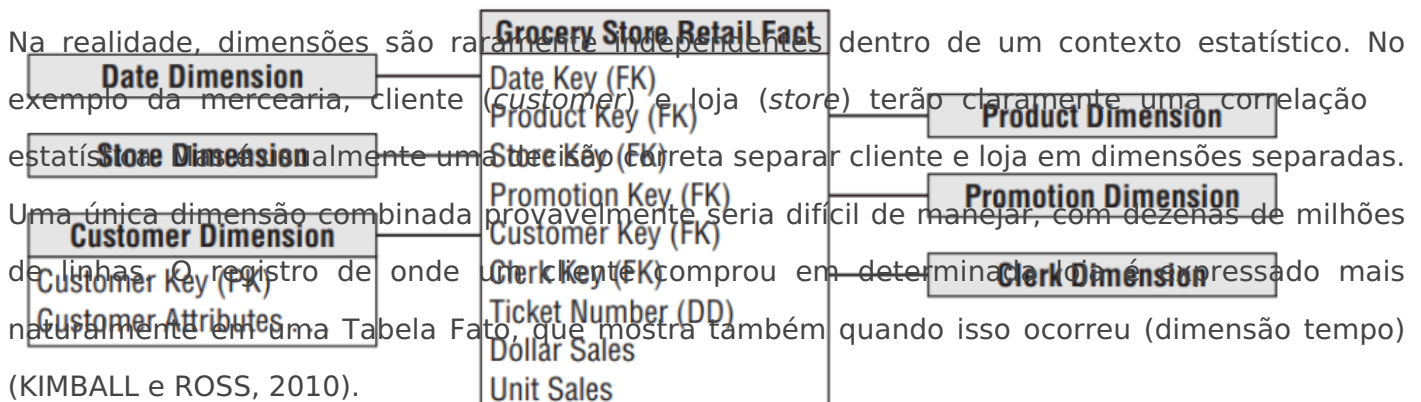
O modelo dimensional deve corresponder à estrutura física dos eventos de captura de dados. Um modelo não deve ser elaborado para atender a um relatório em específico (*report-of-the-day*). Os processos de negócios de uma empresa ultrapassam as fronteiras dos departamentos e funções organizacionais. Em outras palavras, você deve construir uma única tabela de fatos para métricas de vendas atômicas em vez de preencher bancos de dados / tabelas semelhantes, mas ligeiramente diferentes, contendo métricas de vendas para as equipes de vendas, *marketing*, logística e finanças separadamente (KIMBALL e ROSS, 2013).

Medições e Contexto

A modelagem dimensional se inicia dividindo o mundo entre medições e contexto. Medições são usualmente numéricas e tomadas repetidamente. Medições numéricas são *fatos*. Fatos estão sempre cercados geralmente por contexto textuais, que são verdadeiros no momento em que os fatos são gerados. Fatos são constituídos de atributos numéricos, muito específicos e bem definidos. Em contraste, o contexto que cerca os fatos é aberto e verboso (KIMBALL e ROSS, 2010).

Embora possamos aglomerar todo o contexto juntamente com cada medida em um vasto e único registro lógico, veremos que, geralmente, é mais conveniente separar o contexto em grupos independentes. Quando armazenados fatos -- vendas em reais de uma mercearia referentes a um produto específico, por exemplo -- naturalmente dividiremos o contexto entre grupos

denominados Nome do Produto, Loja, Horário, Cliente, Balconista, dentre outros. Chamamos esses agrupamentos lógicos de *dimensões*, e assumimos informalmente que essas dimensões são independentes. A figura abaixo mostra um modelo dimensional para um armazenamento típico de fatos para uma mercearia (KIMBALL e ROSS, 2010).



Assumir a independência de dimensões significa que todas as dimensões, como produto, loja e cliente são independentes do fator tempo. Mas devemos levar em conta a mudança lenta e esporádica dessas dimensões ao longo do tempo. De fato, como mantenedores do *Data Warehouse*, assumimos o compromisso de representar essas mudanças. Esta situação dá origem à técnica denominada *Slowly Changing Dimensions* (SCD) (KIMBALL e ROSS, 2010).

Relacionando os dois Mundos da Modelagem

Modelos dimensionais são modelos relacionais, em que Tabelas Fato estão na terceira forma normal e as Tabelas de Dimensão estão na segunda forma normal, confusamente chamadas de desnormalizadas. Lembre-se que a grande diferença entre a segunda e a terceira forma normal é que os valores repetidos (redundantes) são removidos da segunda forma normal por meio da criação de novas tabelas (*snowflake*). O fato de removermos os valores de dimensão da Tabela Fato, e colocando esses valores em suas próprias tabelas, coloca a Tabela Fato na terceira forma normal (KIMBALL e ROSS, 2010).

Devemos resistir ao impulso de colocar as tabelas de dimensão na terceira forma normal (*snowflake*), pois tabelas únicas (*flat*) são muito mais eficientes para recuperação de dados por meio de instruções SQL. Em particular, os atributos de dimensão com muitos valores repetidos são alvos perfeitos para índices de bitmap. Colocar uma dimensão na terceira forma normal (*snowflaking*), embora não seja incorreto, elimina a possibilidade de utilização de índices de bitmap e aumenta a percepção de complexidade no design (KIMBALL e ROSS, 2010).

Lembre-se que na área de apresentação de um DW não precisamos nos preocupar em forçar a aplicação de regras de dados muito rígidas, exigindo dimensões *snowflaked* (terceira forma normal). A garantia de integridade já é garantida no estágio de ETL (*staging ETL system*) (KIMBALL e

ROSS, 2010).

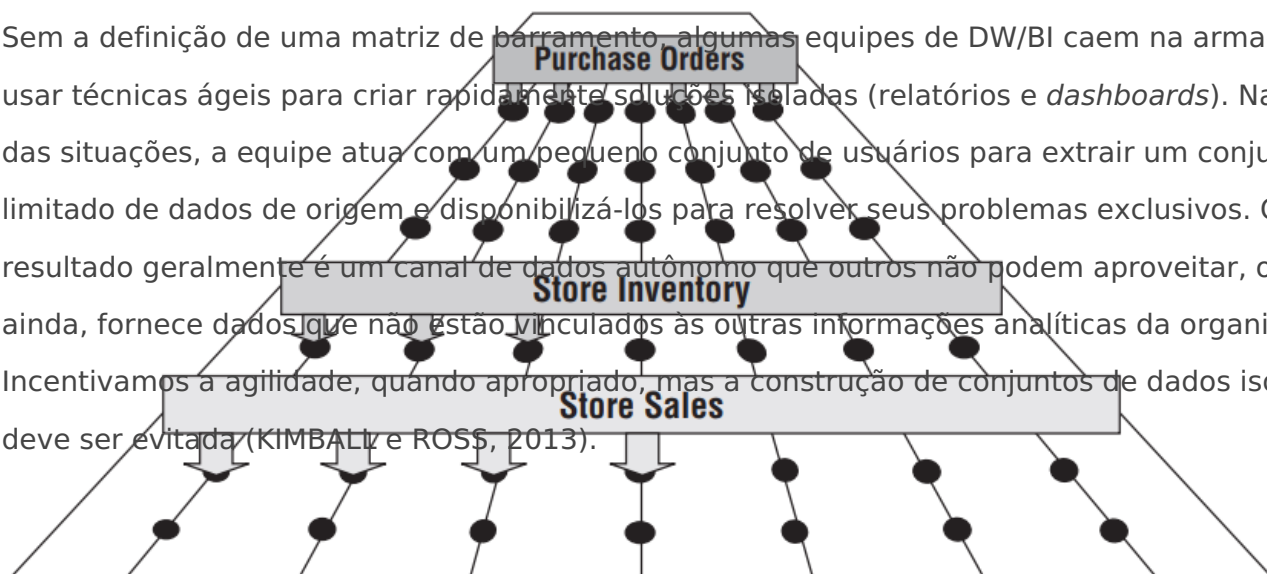
Arquitetura de Barramento do Enterprise Data Warehouse

A arquitetura de barramento do *Enterprise Data Warehouse* (*Enterprise Data Warehouse Bus Architecture*) fornece uma abordagem incremental para construir o sistema de DW/BI corporativo. Essa arquitetura decompõe o processo de planejamento de DW/BI em partes gerenciáveis, concentrando-se nos processos de negócios, enquanto oferece integração por meio de dimensões padronizadas que são reutilizadas em todos os processos. Ela fornece uma estrutura de arquitetura, ao mesmo tempo em que decompõe o programa para incentivar implementações ágeis gerenciáveis correspondentes às linhas na matriz de barramento do EDW (KIMBALL e ROSS, 2013).

Ao definir uma interface de barramento padrão para o ambiente DW/BI, modelos dimensionais separados podem ser implementados por diferentes grupos em momentos diferentes. Deste modo, as áreas de negócio distintas podem se conectar e coexistirem de maneira complementar.

Como ilustrado pelo diagrama abaixo, podemos vislumbrar diversos processos de negócio plugados em um barramento comum. Cada processo da cadeia de valor de uma organização pode exigir a criação de uma família de modelos dimensionais, que compartilham uma série de dimensões conformadas.

Sem a definição de uma matriz de barramento, algumas equipes de DW/BI caem na armadilha de usar técnicas ágeis para criar rapidamente soluções isoladas (relatórios e *dashboards*). Na maioria das situações, a equipe atua com um pequeno conjunto de usuários para extrair um conjunto limitado de dados de origem e disponibilizá-los para resolver seus problemas exclusivos. O resultado geralmente é um canal de dados autônomo que outros não podem aproveitar, ou pior ainda, fornece dados que não estão vinculados às outras informações analíticas da organização. Incentivamos a agilidade, quando apropriado, mas a construção de conjuntos de dados isolados deve ser evitada (KIMBALL e ROSS, 2013).



Matriz de Barramento

A matriz de barramento do Enterprise Data Warehouse (*Enterprise Data Warehouse Bus Matrix*) é uma ferramenta essencial para projetar e comunicar a arquitetura de barramento do EDW. Como pode ser visto na imagem abaixo, as linhas da matriz são processos de negócios e as colunas são dimensões. As células marcadas da matriz indicam se uma dimensão está associada a um determinado processo de negócios. A equipe de design verifica cada linha para testar se uma dimensão candidata está bem definida para o processo de negócios e também verifica cada coluna para ver onde uma dimensão deve ser conformada em vários processos de negócios. Além das considerações de projeto técnico, a matriz de barramento é usada como entrada para priorizar projetos de DW/BI com gestão de negócios, pois as equipes devem implementar uma linha da matriz por vez (KIMBALL e ROSS, 2013).

Tabela Fato

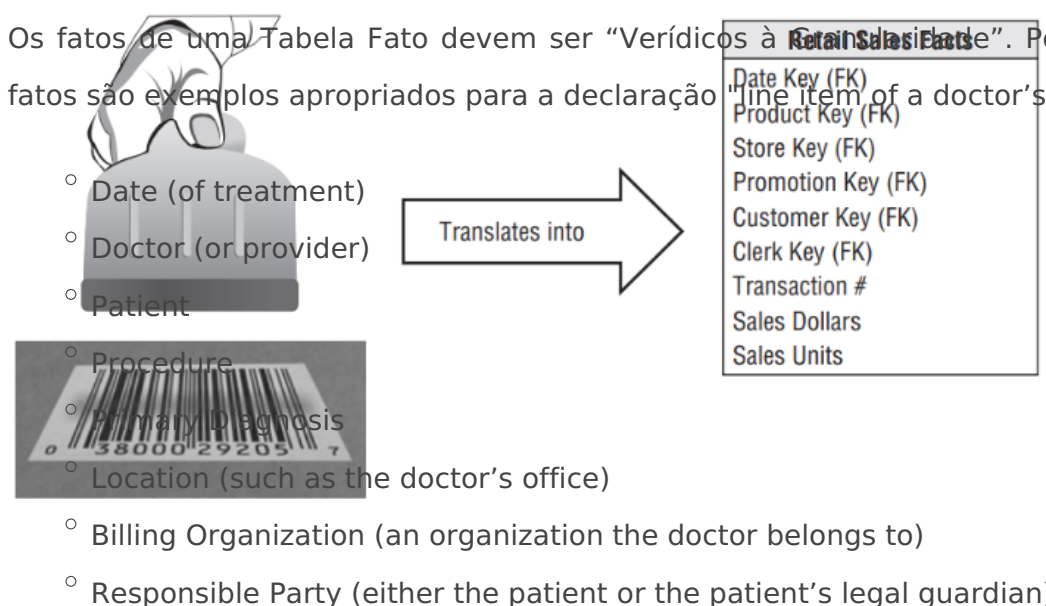
Tabelas Fato são tabelas para armazenamento de medidas. A maioria das medidas armazenadas em Tabelas Fato dizem respeito a séries temporais, em que são armazenados timestamps e chaves estrangeiras conectando a dimensões de data calendário.

Cada linha em uma Tabela Fato corresponde a um evento de medição. A ideia de que um evento de medição no mundo físico tem um relacionamento de um para um com uma única linha na tabela de fatos correspondente é um princípio fundamental para a modelagem dimensional. Todo o resto é construído a partir dessa premissa (KIMBALL e ROSS, 2013).

Tabelas Fato são a figura central de qualquer modelo dimensional. Toda Tabela Fato deve possuir uma única e explícita granularidade definida. A granularidade de uma Tabela Fato deve ser definida durante o processo de modelagem. Kimball e Ross (2010) recomendam que a definição da granularidade (explicitar exatamente o que uma Tabela Fato representa) seja a primeira ação a ser realizada na modelagem de uma Tabela Fato. Podem ser feitas declarações atômicas ou de alto nível, representando agrupamentos e sumarizações. Após definir a granularidade, deve-se então definir precisamente as dimensões possíveis associadas.

Kimball e Ross (2010) recomendam que os detalhes mais atômicos disponíveis nos sistemas de origem devem ser refletidos no DW por meio de Tabelas Fato. Um registro fato (fact record) em um modelo dimensional é criado como uma resposta 1:1 para a medida de um evento em um processo de negócio específico. Tabelas Fato são definidas pela física do mundo real.

Os fatos de uma Tabela Fato devem ser “Verídicos à Realidade”. Por exemplo, os seguintes fatos são exemplos apropriados para a declaração “Line item of a doctor’s bill”:



- Primary Payer (often an insurance plan)
- Secondary Payer (maybe the responsible party's spouse's insurance plan)

Outros fatos, como “Valor cobrado no ano até a data corrente para o paciente para todos os tratamentos” não são Verídicos. Neste caso, quando aplicações de BI combinam registros de fato arbitrariamente, estes Fatos Inverídicos produzem resultados sem sentido e inúteis. Olhando desta maneira, este tipo de fato é perigoso pois induz usuários de negócio a erros. Este tipo de fato agregado deve ser omitido do design do DW e calculado diretamente na aplicação de BI (KIMBALL e ROSS, 2013).

Como os dados de medição são esmagadoramente o maior conjunto de dados, eles não devem ser replicados em vários locais para várias funções organizacionais na organização. Permitir que usuários de negócios de várias organizações acessem um único repositório centralizado para cada conjunto de dados garante o uso de dados consistentes em toda a empresa (KIMBALL e ROSS, 2013).

É teoricamente possível que um fato medido seja textual; no entanto, a condição raramente surge. Na maioria dos casos, uma medida textual é uma descrição de algo e é extraída de uma lista discreta de valores. O designer deve fazer todos os esforços para colocar os dados textuais em dimensões onde possam ser correlacionados de forma mais eficaz com os outros atributos de dimensão textual. Você não deve armazenar informações textuais redundantes em tabelas de fatos. A menos que o texto seja exclusivo para cada linha na tabela de fatos, ele pertence à tabela de dimensão. Um fato de texto verdadeiro é raro porque o conteúdo imprevisível de um fato de texto, como um comentário de texto de forma livre, torna-o quase impossível de ser analisado (KIMBALL e ROSS, 2013).

Tabelas Fato correspondem a medidas resultantes de eventos observáveis, e não a demandas de relatórios específicos (KIMBALL e ROSS, 2013).

Tabela de Dimensão

Tabelas de dimensão são companheiras integrais de Tabelas Fato. Tabelas de dimensão possuem descrições textuais que descrevem o contexto associado à medida realizada pelo processo de negócio. Elas descrevem o “quem, o quê, onde, quando, como e porque”, associado ao evento. Como ilustrado abaixo, tabelas de dimensão geralmente possuem muitos atributos. Não é incomum que uma tabela de dimensão possua de 50 a 100 atributos, embora algumas possuam apenas uma quantidade reduzida (Kimball e Ross, 2013).

Os atributos de dimensão servem como a fonte primária de restrições em consultas, agrupamentos e rótulos para colunas de relatórios. Por exemplo, quando um usuário deseja ver o valor vendido por marca, a coluna “marca” deve estar disponível como um atributo da dimensão associada (Kimball e Ross, 2013).

Os atributos devem consistir em palavras reais em vez de abreviações. Você deve se esforçar para minimizar o uso de códigos em tabelas de dimensão, substituindo-os por atributos textuais mais detalhados. Às vezes, os códigos ou identificadores operacionais têm um significado legítimo de negócio para os usuários, ou são requeridos ao se comunicar com a área operacional. Nestes casos, os códigos devem aparecer como atributos de dimensão explícitos, além dos atributos textuais de fácil interpretação. Códigos operacionais às vezes têm inteligência embutida. Por exemplo, os primeiros dois dígitos podem identificar a linha de negócios, enquanto os próximos dois dígitos podem identificar a região global. Em vez de forçar os usuários a interrogar ou filtrar em substrings dentro dos códigos operacionais, extraia os significados incorporados e apresente-os aos usuários como atributos de dimensão separados que podem ser facilmente filtrados, agrupados e reportados (Kimball e Ross, 2013).

Kimball e Ross (2013) mostram que quanto mais granular é o dado, maior a sua dimensionalidade. Deste modo, o dado deve ser armazenado em um DW em sua menor granularidade e maior número de dimensões possíveis. Isso deve ser feito para que o usuário de negócio tenha a maior gama possível de possibilidades de filtros e agrupamentos na elaboração de relatórios ad-hoc.

Product Dimension
Product Key (PK)
SKU Number (Natural Key)
Product Description
Brand Name
Category Name
Department Name
Package Type
Package Size
Abrasive Indicator
Weight
Weight Unit of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
...

Embora seja aceitável, relacionamentos entre dimensões devem ser evitados. Na maioria dos casos, as correlações entre dimensões (*Outrigger Dimensions*) devem ser denotadas por meio de Tabelas Fato (Kimball e Ross, 2013).

Quando as dimensões são separadas, alguns designers desejam criar uma pequena tabela com apenas as duas chaves de dimensão para mostrar a correlação sem usar a tabela de fato. Em muitos cenários, essa tabela bidimensional é desnecessária. Não há razão para evitar a tabela fato para responder a esta consulta de relacionamento. As tabelas fato são incrivelmente eficientes porque contêm apenas chaves de dimensão e medições, juntamente com a dimensão degenerada ocasional. A tabela fato é criada especificamente para representar as correlações e os relacionamentos muitos para muitos entre as dimensões (Kimball e Ross, 2013).

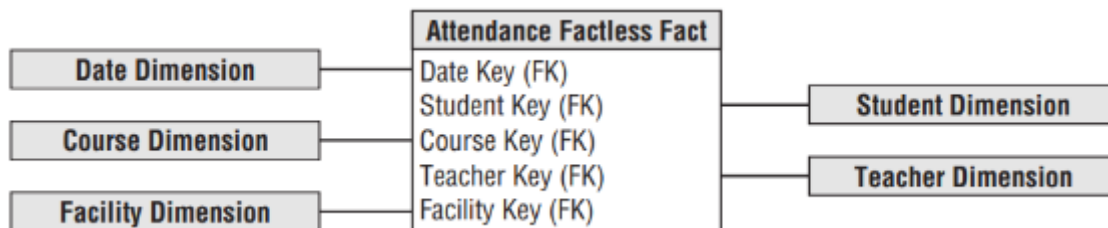
Star Schema

Agora que você entende as Tabelas Fato e Dimensões, é hora de reunir os blocos de construção em um modelo dimensional, conforme mostrado na figura abaixo. Cada processo de negócio é representado por um modelo dimensional que consiste em uma Tabela Fato contendo os dados numéricos do evento medido, cercada por um grupo de tabelas de dimensão, com o contexto que era verdadeiro no momento em que o evento ocorreu. Essa estrutura característica em forma de estrela costuma ser chamada de junção em estrela (*Star Join*) (KIMBALL e ROSS, 2013).



Tabela Fato sem Fato

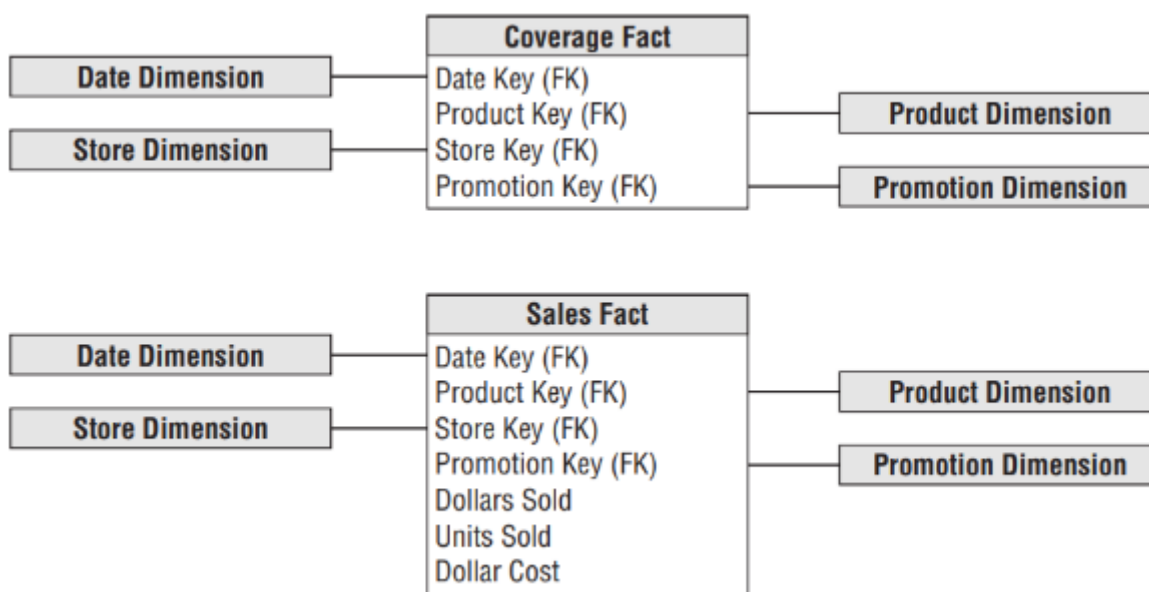
Tabela Fato sem fato (*Factless Fact Tables*) são Tabelas Fato que não possuem nada além de chaves estrangeiras para dimensões. O primeiro tipo de *Factless Table* diz respeito a uma tabela que armazena eventos. Diversas Tabelas Fato para rastreamento de eventos tornam-se Tabelas Fato sem fato. Um exemplo seria uma Tabela Fato para armazenar atendimentos a estudantes em uma faculdade, como pode ser visto na figura abaixo:



Esse tipo de tabela torna-se uma Tabela Fato sem fato, pois não existem fatos óbvios gerados a cada atendimento de um estudante. Devido a ausência de fatos, e no intuito de melhorar a legibilidade de instruções SQL sobre esse tipo de tabela, alguns engenheiros de dados optam por incluir um fato fictício ao final da tabela, com valor constante 1, por exemplo, “atendimento”.

Um segundo tipo de Tabela Fato sem fato é denominado tabela de cobertura (*coverage table*).

Uma tabela de cobertura típica é mostrada abaixo:



Tabelas de cobertura são frequentemente necessárias quando as Tabelas Fato principais são esparsas. A figura acima mostra também uma tabela de vendas (*Sales Fact*) que armazena as vendas de produtos em lojas em dias particulares, sob condições diversas de promoção. A Tabela Fato principal responde diversas perguntas de negócio interessantes, mas não pode dizer nada sobre coisas que não ocorreram. Por exemplo, não pode responder perguntas como: “Quais produtos estavam em promoção, mas não venderam?”, pois a tabela possui apenas produtos que foram vendidos (Kimball e Ross, 2013).

Tabela Fato de Snapshot Periódico

Uma linha em uma Tabela Fato de Snapshot Periódico (*Periodic Snapshot Fact Table*) sumariza diversas medidas de eventos ocorridos sobre um período pré-definido, como um dia, uma semana, ou um mês. Neste caso, o grão é o período, não a transação individual. Este tipo de tabela possui muitos fatos porque qualquer evento de medição consistente com a granulação da tabela de fatos é permitido. São tabelas geralmente uniformemente densas em suas chaves estrangeiras, pois mesmo que nenhuma atividade tenha ocorrido no período, uma linha é tipicamente inserida para cada fato, contendo zero ou *null* (Kimball e Ross, 2013).

Fatos Conformados

Se a mesma medida aparece em diferentes Tabelas Fato, deve ser tomado o devido cuidado para que as definições técnicas dos fatos sejam idênticas. Se fatos separados são consistentes, tais fatos conformados devem possuir a mesma nomenclatura. Mas se os fatos são incompatíveis, eles devem possuir nomenclaturas diferentes, que servirão como alerta aos usuários de negócio (Kimball e Ross, 2013).

Chaves Dimensionais

Se fatos são medidas reais geradas rapidamente, concluímos que Tabelas Fato criam uma situação de relacionamento muitos-para-muitos (M:N) entre as dimensões. Muitos clientes compram muitos produtos em diversas lojas em diversos momentos (KIMBALL e ROSS, 2010).

Deste modo, modelamos as medidas como Tabelas Fato com diversas chaves estrangeiras (*Foreign Key* - FK), referenciando as entidades de contexto (dimensões). E as tabelas de dimensões (entidades de contexto) possuem, cada uma, a sua chave primária (*Primary Key* - PK) (KIMBALL e ROSS, 2010).

Uma Tabela Fato em um modelo dimensional *Star Schema* é constituída, além das medidas, por múltiplas FKs, cada uma pareada com a chave primária em uma dimensão. Note que este design possibilita às tabelas de dimensão possuírem chaves primárias não existentes na Tabela Fato. Esta situação é consistente com a manutenção da integridade referencial e apropriada para a modelagem dimensional (KIMBALL e ROSS, 2010).

No mundo real, existem diversas razões convincentes para a construção de pares FK-PK como chaves substitutas (*Surrogate Keys*), que são apenas números inteiros sequenciais. É um grande erro definir chaves em um DW com base em chaves naturais (*Natural Keys*), provenientes das fontes operacionais (KIMBALL e ROSS, 2010).

Ocasionalmente, uma medida perfeitamente válida poderá envolver alguma dimensão ausente. Talvez, em algumas situações, um produto poderá ser vendido a um cliente por meio de uma transação sem uma loja definida. Neste caso, ao invés de armazenarmos *null* na FK de loja, podemos manter um registro especial na dimensão loja, com o valor “No Store”, e utilizá-lo para esta finalidade, por exemplo (KIMBALL e ROSS, 2010).

Logicamente, uma Tabela Fato não precisa de uma chave primária. Dependendo da situação, duas observações legítimas poderão ser representadas de forma idêntica na Tabela Fato. Da perspectiva prática, esta é uma ideia terrível, pois teremos muita dificuldade em separar tais registros por meio da utilização de instruções SQL. Será difícil verificar a qualidade dos dados se diversos registros forem representados de forma idêntica em nossa Tabela Fato (KIMBALL e ROSS, 2010).

Tabela Fato Transacional

De acordo com Kimball e Ross (2013) uma Tabela Fato Transacional (*Transaction Fact Table*) é um tipo de tabela fato onde são mantidos os fatos na menor granularidade possível. Neste caso, é mantida uma transação para cada evento real. Uma Tabela Fato é do tipo transacional se:

- Cada evento é armazenado apenas uma vez;
- Há uma coluna de data indicando quando aquele evento ocorreu;
- Há um atributo identificador, que identifica cada evento unicamente.

A quantidade de registros da tabela é a mesma da origem.

Tabela Fato de Snapshot Acumulado

De acordo com (Kimball e Ross, 2013), uma Tabela Fato de Snapshot Acumulado (*Accumulating Snapshot Fact Table*) é menos frequente que as demais e corresponde a um processo previsível, cujos início e fim são bem definidos. Capturam os resultados de eventos-chave em uma relação de processos relacionados. As linhas da tabela são carregadas quando o primeiro evento ocorre, ou quando um marco histórico é alcançado. Diferentemente de outros tipos de Tabelas Fato, neste caso, linhas pré-existentes são atualizadas para refletir resultados correntes, ou acumulados de cada evento.

- Talvez usuários de negócio queiram analisar o seu pipeline de aquisições. Eles já possuem fatos atômicos individuais que capturam os detalhes ricos associados a cada evento de transação no pipeline, como a submissão de requisições de compra, emissão de pedidos de compra, recebimento de entregas, recebimento de faturas, emissão de pagamentos, etc. Embora cada um desses eventos compartilhem dimensões comuns, como Produto, Vendedor e Solicitante, as Tabelas Fato individuais possuem dimensões e métricas únicas.
- Suponha que algum usuário de negócio queira saber o quão rápido ordens de compra são emitidas após a submissão de requisições. Ou qual a discrepância entre a quantidade pedida versus o que foi recebido. Ou qual a duração ou tempo de atraso entre o recibo das faturas e o pagamento. *Accumulating Snapshot Fact Table* vêm para auxiliar neste tipo de análise.
- *Accumulating Snapshot Fact Table* é um tipo de Tabela Fato consolidada, que facilita a busca por determinado tipo de informação.

Natural Key

Uma *Natural Key* (NK) é uma chave única, composta por uma ou mais colunas, que identificam um registro unicamente, possuindo significado para o negócio. Uma NK possui relação semântica com os demais atributos em uma relação. Ex. CPF, CEP, RG, etc.

Dimensões de Etapa (Step Dimensions)

Processos sequenciais, como eventos de página da *Web*, normalmente têm uma linha separada em uma tabela de fatos para cada etapa de um processo. Para saber onde a etapa individual se encaixa na sessão geral, é usada uma dimensão de etapa que mostra qual número de etapa é representado pela etapa atual e quantas etapas a mais foram necessárias para concluir a sessão (Kimball e Ross, 2013).

Tabelas Fato Agregadas

Em adição às Tabelas Fato que armazenam dados atômicos relativos a fatos únicos de processos, tabelas agregadas são muitas vezes criadas. Essas tabelas combinam métricas de múltiplos processos em um nível comum de detalhe. São tabelas complementares às Tabelas Fato detalhadas, e não substitutas (KIMBALL e ROSS, 2013).

Agregações necessariamente agrupam e/ou removem dimensões presentes em Tabelas Fato atômicas. Deste modo, agregações sempre devem ser utilizadas juntamente com os dados atômicos, visto que estes possuem dimensões mais detalhadas.

Fatos Aditivos

No coração de qualquer Tabela Fato existe uma lista de fatos que representam medidas. Como a maioria das Tabelas Fato são grandes, com milhões ou até bilhões de registros, dificilmente utilizaremos um único registro para responder a alguma questão de negócio. Ao invés disso, geralmente recuperamos uma grande quantidade de registros, comprimidos em um formato amigável de soma (adição), contagem, média, valor máximo, valor mínimo, etc. Mas por questões práticas, o formato mais utilizado, de longe, é a soma. Deste modo, sempre que possível, devemos armazenar fatos aditivos (que permitam a soma). Deste modo, no exemplo da mercearia, não precisamos armazenar o preço unitário de um produto. Pois conseguiremos este valor simplesmente dividindo o valor da compra (*dollar sales*) pela quantidade comprada (*unit sales*) (KIMBALL e ROSS, 2010).

Alguns fatos, como saldos bancários e níveis de estoque, representam intensidades difíceis de expressar em um formato aditivo. Neste caso, tratamos este tipo de fato semi-aditivo como se fossem aditivos, mas antes de apresentar a informação para usuários de negócio, dividimos a resposta pela quantidade de períodos para termos o resultado correto. Esta técnica é chamada de média sobre o tempo (*averaging over time*) (KIMBALL e ROSS, 2010).

Algumas Tabelas Fato representam medições de eventos sem fato, esse tipo de tabela é chamada Tabela Fato sem Fato (*factless fact tables*). O exemplo clássico é o registro de atendimentos de estudantes de uma determinada turma para um dia em particular. As dimensões podem ser “dia”, “estudante”, “professor”, “curso” e “local”, mas veja que não há uma métrica claramente definida (KIMBALL e ROSS, 2010).

Dimensões Multivaloradas e Tabelas Ponte (bridge table)

Em um esquema dimensional clássico, cada dimensão anexada a uma tabela de fatos tem um único valor consistente com a granulação da tabela de fatos. Mas há uma série de situações em que uma dimensão é legitimamente multivalorada. Por exemplo, um paciente que recebe um tratamento de saúde pode ter vários diagnósticos simultâneos. Nesses casos, a dimensão multivalorada deve ser anexada à tabela de fatos por meio de uma chave de dimensão de grupo para uma tabela de ponte (*bridge table*) com uma linha para cada diagnóstico simultâneo em um grupo (Kimball e Ross, 2013).

Tabelas Fato Consolidadas

Frequentemente, é conveniente combinar fatos de vários processos em uma única Tabela Fatos consolidada se eles puderem ser expressos na mesma granulação. Por exemplo, dados de vendas podem ser consolidados com previsões de vendas em uma única Tabela Fatos para simplificar a tarefa de analisar os valores reais e os previstos, em comparação com a realização de uma operação de *drill-across*, usando Tabelas Fato separadas. Tabelas Fatos consolidadas aumentam a carga do processamento de ETL, mas facilitam a carga analítica nos aplicativos de BI. Esse tipo de Tabela Fato deve ser considerada para métricas frequentemente analisadas em conjunto (KIMBALL e ROSS, 2013).

Quando os fatos de vários processos de negócios são combinados em uma tabela de fatos consolidada, eles devem estar no mesmo nível de granularidade e dimensionalidade. Como os fatos separados raramente vivem naturalmente em um grão comum, você é forçado a eliminar ou agregar algumas dimensões para suportar a correspondência um-para-um, mantendo os dados atômicos em tabelas de fatos separadas. As equipes de projeto não devem criar fatos ou dimensões artificiais na tentativa de forçar a consolidação de dados de fatos com granulação diferente (KIMBALL e ROSS, 2013).

Tabelas Fato consolidadas contêm chaves estrangeiras para dimensões conformadas reduzidas, bem como fatos agregados criados pela soma de medidas de tabelas fato mais atômicas (KIMBALL e ROSS, 2013).

Fatos Agregados e Atributos de Dimensões

Usuários de negócio geralmente estão interessados em restringir a dimensão do cliente com base em métricas de desempenho agregadas, como filtrar todos os clientes que gastaram mais de um determinado valor no ano passado ou talvez durante a vida do cliente. Fatos agregados podem ser colocados em uma dimensão para restrição e como rótulos de linha para relatórios. As métricas geralmente são apresentadas como intervalos em faixas na tabela de dimensões. Os atributos de dimensão que representam as métricas de desempenho agregadas adicionam carga ao processamento de ETL, mas aliviam a carga analítica na camada de BI (Kimball e Ross, 2013).

Dimensões Genéricas Abstratas

Alguns modeladores são atraídos por dimensões genéricas abstratas. Por exemplo, seus esquemas incluem uma única dimensão de localização genérica em vez de atributos geográficos incorporados nas dimensões de loja, depósito e cliente. Da mesma forma, sua dimensão de pessoa inclui linhas para funcionários, clientes e contatos de fornecedores porque todos são seres humanos, independentemente de serem significativamente diferentes devido aos atributos coletados para cada tipo. Dimensões genéricas abstratas devem ser evitadas em modelos dimensionais.

Os conjuntos de atributos associados a cada tipo geralmente diferem. Se os atributos forem comuns, como um estado geográfico, eles devem ser rotulados de forma exclusiva para distinguir o estado de uma loja do estado de um cliente. Finalmente, despejar todas as variedades de locais, pessoas ou produtos em uma única dimensão invariavelmente resulta em uma tabela de dimensão maior. A abstração de dados pode ser apropriada no sistema de origem operacional ou no processamento ETL, mas afeta negativamente o desempenho e a legibilidade da consulta no modelo dimensional (Kimball e Ross, 2013).

Dimensões de comentário

Em vez de tratar os comentários de forma livre como métricas textuais em uma tabela de fatos, eles devem ser armazenados fora da tabela de fatos em uma dimensão de comentários separada (ou como atributos em uma dimensão com uma linha por transação se a cardinalidade dos comentários corresponder ao número de transações exclusivas) com uma chave estrangeira correspondente na tabela de fatos (Kimball e Ross, 2013).

Dimensões de Auditoria

Quando uma linha da tabela de fatos é criada pelo processo de ETL, é útil criar uma dimensão de auditoria contendo os metadados de processamento de ETL conhecidos no momento. Uma linha de dimensão de auditoria simples pode conter um ou mais indicadores básicos de qualidade de dados, talvez derivados do exame de um esquema de evento de erro que registra violações de qualidade de dados encontrados durante o processamento dos dados. Outros atributos de dimensão de auditoria úteis podem incluir variáveis de ambiente que descrevem as versões do código ETL usadas para criar as linhas de fatos ou os carimbos de tempo de execução do processo ETL (KIMBALL e ROSS, 2013).

Essas variáveis de ambiente são especialmente úteis para fins de conformidade e auditoria porque permitem que as ferramentas de BI façam uma busca detalhada para determinar quais linhas foram criadas com quais versões do software ETL.

Dimensão Conformada

Uma Dimensão Conformada (também chamada de Dimensão Compartilhada ou Dimensão Mestre) é uma dimensão que possui o mesmo significado para todas as Tabelas Fato que podem fazer junção à mesma. Uma das maiores responsabilidades em manter o DW está em estabelecer, publicar, manter e fazer valer as dimensões conformadas (KIMBALL e ROSS, 2010).

Tabelas de dimensão são conformes quando atributos com o mesmo nome em tabelas diferentes possuem o mesmo significado e conteúdo. Informações de diferentes Tabelas Fato podem ser combinadas em um único relatório por meio de atributos de dimensões conformadas. Quando um atributo conformado é utilizado como um label (isto é, no GROUP BY da instrução SQL), os resultados de diferentes Tabelas Fato podem ser alinhados na mesma linha em um relatório drill-across (KIMBALL e ROSS, 2013).

O estabelecimento de uma dimensão conformada é um passo muito importante. Uma dimensão conformada de Clientes, por exemplo, é uma tabela mestre de clientes com uma chave e atributos bem definidos e íntegros. É provável que a dimensão conformada de clientes seja aglomerado de dados de vários sistemas legados e possivelmente de fontes externas. O campo de endereço, por exemplo, deve ser constituído do endereço mais completo e atualizado possível para cada cliente (KIMBALL e ROSS, 2010).

A dimensão conformada de produtos, por exemplo, é a lista mestre de todos os produtos comercializados pela empresa, incluindo todos os atributos do produto e todas as suas agregações, como categoria e departamento. Uma boa dimensão de produtos, assim como uma boa dimensão de clientes, deve ter ao menos 50 atributos textuais (KIMBALL e ROSS, 2010).

Idealmente, a dimensão conformada de localização deve ser baseada em pontos específicos do mapa, como endereços específicos de ruas ou até mesmo latitudes e longitudes. Pontos específicos no espaço se acumulam em todas as hierarquias geográficas concebíveis, incluindo cidade-região-estado-país, bem como códigos postais, territórios e regiões de vendas (KIMBALL e ROSS, 2010).

A dimensão conformada de data quase sempre será uma tabela de dias individuais, abrangendo uma década ou mais. Cada dia terá muitos atributos úteis extraídos dos calendários legais dos vários estados e países com os quais a empresa lida, bem como períodos fiscais especiais e

temporadas de marketing relevantes (KIMBALL e ROSS, 2010).

Dimensões conformadas são extremamente importantes para o DW. Sem uma adesão estrita às dimensões conformadas, o DW não pode funcionar como um todo integrado.

Dimensões reduzidas (Shrunken Dimensions)

Dimensões reduzidas são dimensões conformadas que são um subconjunto de linhas e/ou colunas de uma dimensão conformada original. São necessárias ao construir Tabelas Fato agregadas. Elas também são necessárias para processos de negócios que capturam dados naturalmente em um nível mais alto de granularidade, como uma previsão por mês e marca (em vez da data mais atômica e do produto associado aos dados de vendas). Outro caso de subconjunto de dimensão conformada ocorre quando duas dimensões estão no mesmo nível de detalhe, mas uma representa apenas um subconjunto de linhas (Kimball e Ross, 2013).

Slowly Changing Dimensions

Slowly Changing Dimensions (SCD) consiste em técnicas para gerenciamento da história de dados dimensionais em um DW. Deve-se ter em mente que as dimensões associadas às Tabelas Fato são afetadas com o passar do tempo. A mudança nos valores de dimensões pode ocorrer às vezes, em alguns casos, meramente para a correção de erros. Mas, às vezes, uma mudança em alguma descrição representa uma mudança verdadeira ocorrida em algum ponto do tempo, como a mudança do nome de um produto, ou de um cliente, por exemplo. Como tais mudanças ocorrem de forma inesperada, e esporadicamente, e com bem menos frequência que medidas em Tabelas Fato, o tópico é denominado *Slowly Changing Dimensions* (SCDs) (Kimball e Ross, 2013).

Os representantes de governança e administração de dados da empresa devem estar ativamente envolvidos nas decisões sobre o tratamento de mudanças em valores de dimensões; A TI não deve fazer determinações por conta própria (Kimball e Ross, 2013).

Kimball e Ross (2013) afirmam que são necessárias apenas três respostas básicas em caso de mudança de descrição em algum atributo de dimensão:

- Tipo 1: Substituição (*Overwrite*):
 - Destruição da história. No exemplo acima, relatórios que filtram ou agrupam pelo campo “cidade” serão afetados. Usuários de negócio precisam ser avisados de que isso pode ocorrer. O DW precisa de uma política explícita e visível de utilização do Tipo 1, como por exemplo: “Erros serão corrigidos”, e/ou, “Não mantemos a história deste campo, mesmo se o seu valor for alterado”.
 - Propagação da mudança. Todas as cópias distribuídas que dependem da dimensão deverão replicar as mudanças simultaneamente.
 - Suponha que em um dia qualquer, o atributo “cidade” de Ralph Kimball, na Tabela Dimensão “empregado”, foi alterado de “Santa Cruz” para “Boulder Creek”. Mais tarde, somos avisados que isso se tratava de uma correção de erro, e não de uma mudança de localização. Neste caso, decidimos por atualizar o valor de “cidade”. Este é um exemplo clássico de uma mudança de Tipo 1. Adequado para casos de correção de erros e para situações em que a mudança não afete a história.

- Embora o Tipo 1 seja o tratamento mais simples e direto para lidar com alterações em valores de dimensões, os seguintes pontos devem ser observados:
- Tipo 2: Adição de um novo registro de dimensão:
 - O Tipo 2 requer que generalizemos a chave primária de “empregado”. Neste caso, é recomendado que a chave primária da relação “empregado” seja algo desconectado da chave primária da relação original, proveniente do sistema relacional. Indica-se, por exemplo, a criação de uma chave primária sequencial. Este tipo de chave primária é chamado *surrogate key* (chave substituta).
 - Vamos alterar o cenário anterior e supor que Ralph Kimball de fato se mudou para uma nova cidade em 18/07/2020. Digamos que a política da organização consiste em manter no DW o histórico exato de endereço residencial de cada funcionário. Esta então seria uma situação clássica para aplicação do Tipo 2.
 - O Tipo 2 exige que seja incluído um novo registro na relação “empregado”, referente ao novo valor para o campo “cidade”. Este tratamento resulta nos seguintes efeitos:
- Tipo 3: Adição de um novo atributo.

- Embora os tipos 1 e 2 vistos anteriormente sejam as técnicas principais utilizadas para tratamento de mudanças em dimensões, precisaremos de uma terceira técnica para abarcar situações em que precisamos alternar entre realidades distintas. Diferentemente de atributos físicos que podem ter apenas um único valor em determinado momento, alguns atributos atribuídos pelo usuário podem legitimamente ter mais de um valor por vez, dependendo do ponto de vista do observador. Por exemplo, uma categoria de produto pode ter mais de uma interpretação. Em uma papelaria, uma caneta de marcação pode ser assimilada na categoria de bens domésticos, ou na categoria de suprimentos artísticos. Usuários e aplicações de BI precisam ter a possibilidade de escolha, em tempo de execução, de qual realidade será aplicada naquele momento.
- O requisito por uma realidade alternativa de visualização de um atributo de dimensão é geralmente acompanhado do requisito de se permitir a separação de versões da realidade nos dados passados e futuros, mesmo que a requisição da inclusão dessa nova realidade tenha chegado ao DW no momento presente.
- Tais requisitos podem ser alcançados por meio da adição de uma nova coluna para cada realidade alternativa do atributo de dimensão. Com esta abordagem, usuários e aplicações de BI podem, a qualquer momento, alterar a realidade por meio da seleção do campo criado para a realidade desejada.

É comum uma mesma dimensão possuir estratégias de tratamento de mudanças específicas para cada atributo (Kimball e Ross, 2013).

Dimensões Degeneradas

Em diversas situações onde a granularidade é um filho de algum evento maior, a chave natural deste evento pai acaba tornando-se um órfão durante o design. No exemplo dado pela Figura 1, a granularidade definida é uma linha da nota de venda (*ticket*). O número do ticket (*ticket number*) é a chave natural da nota de venda. Como os dados do ticket foram espalhados em dimensões de contexto, o número do ticket foi deixado para trás, sem atributos específicos. Neste caso, o número do *ticket* é incluído diretamente na Tabela Fato. Chamamos esta situação de dimensão degenerada (*degenerate dimension*). O número do *ticket* é útil pois é a cola que segura os itens presentes em uma mesma nota (KIMBAL e ROSS, 2013).

Mantendo a Granularidade na Modelagem Dimensional

Embora teoricamente qualquer mistura de fatos possa ser incluída em uma mesma tabela, um design dimensional apropriado permite apenas fatos na mesma granularidade (mesma dimensionalidade) coexistirem em uma mesma Tabela Fato. Uma granularidade uniforme garante que todas as dimensões serão utilizadas por todos os fatos, reduzindo drasticamente a possibilidade de equívocos durante a combinação de dados de diferentes granularidades. Por exemplo, geralmente não faz sentido adicionar dados diários a dados anuais. Quando você tem fatos em duas granularidades diferentes, coloque-os em tabelas separadas (KIMBALL e ROSS, 2010).

O poder do modelo dimensional vem da sua aderência a uma granularidade específica. Uma definição clara da granularidade de uma Tabela Fato possibilita a definição dos modelos físico e lógico. Uma definição confusa ou imprecisa impõe ameaças em todos os aspectos de design, desde os processos de ETL até as ferramentas de visualização que farão uso do dado (KIMBALL e ROSS, 2010).

A granularidade de uma Tabela Fato é a definição de negócio de uma medida que resulta na criação de um registro fato (*fact record*). Toda a definição de granularidade deve ser iniciada do menor nível possível (atômico), relativo para o evento gerador do fato. Manter aderência à granularidade significa construir Tabelas Fato em torno de cada evento de medição de processos de negócio atômicos. Essas tabelas são simples de serem implementadas, mas proveem uma fundação durável e flexível, capaz de endereçar questões de negócio e relatórios do dia-a-dia (KIMBALL e ROSS, 2010).

Pense Dimensionalmente

Ao reunir os requisitos para uma iniciativa de DW/BI, você precisa ouvir e em seguida, sintetizar as descobertas em torno dos processos de negócios. Quando especificar o escopo de um projeto, mantenha o foco em um único processo de negócio por projeto, evitando a elaboração de *dashboards* com inúmeros processos de negócio associados (KIMBALL e ROSS, 2013).

Embora seja fundamental que a equipe de DW/BI se concentre nos processos de negócios, é igualmente importante manter o alinhamento com a equipe de negócios. Devido às políticas históricas de financiamento da TI, a empresa pode estar mais familiarizada com implantações de soluções a nível departamental. Você precisa mudar a mentalidade dos envolvidos para a implementação de soluções de DW/BI para uma perspectiva de processos de negócios. Felizmente, gestores de negócio geralmente adotam essa abordagem porque reflete seu pensamento sobre os principais indicadores de desempenho. Além disso, eles viveram com inconsistências, debates incessantes e reconciliações sem fim causadas pela abordagem departamental, então eles estão prontos para uma nova abordagem (KIMBALL e ROSS, 2013).

Ao atuar com as lideranças de negócios, classifique cada processo em escala de valor e viabilidade e, em seguida, ataque primeiramente os processos com as pontuações de maior impacto e viabilidade. Embora a priorização seja uma atividade conjunta com o negócio, sua compreensão subjacente dos processos de negócios da organização é essencial para sua eficácia e subsequente capacidade de ação (KIMBALL e ROSS, 2013).

Os programas de governança de dados devem se concentrar primeiro nas dimensões principais. Dependendo do setor da indústria, a lista pode incluir data, cliente, produto, funcionário, aluno, corpo docente e assim por diante. Pensar nos substantivos centrais usados para descrever o negócio se traduz em uma lista de esforços de governança de dados a serem liderados por especialistas no assunto e representantes da comunidade de negócios. Estabelecer responsabilidades de governança de dados para esses substantivos é a chave para implementar dimensões que entreguem consistência e atendam às necessidades de negócios de filtragem analítica, agrupamento e rotulagem. Dimensões robustas se traduzem em sistemas DW/BI robustos (KIMBALL e ROSS, 2013).

Modelos dimensionais devem ser elaborados em colaboração com a área de negócios, e não por indivíduos que não entendem completamente o negócio e suas necessidades (KIMBALL e ROSS, 2013).

Schemas de Eventos de Erros

Gerenciar a qualidade de dados em um Data Warehouse requer um sistema abrangente de qualidade que testa os dados à medida que fluem dos sistemas de origem para a plataforma de BI. Quando o *script* de qualidade de dados detecta um erro, esse evento é registrado em um esquema dimensional especial que está disponível apenas no backend do processo de ETL. Esse esquema consiste em uma tabela de fatos de evento de erro cuja granulação é o evento de erro individual e uma tabela de fatos de detalhes de evento de erro associada cuja granulação é cada coluna em cada tabela que participa de um evento de erro (KIMBALL e ROSS, 2013).

Surrogate Key

Uma *Surrogate Key* (SK), ou chave substituta, assim como uma Natural Key (NK), é um identificador único em uma tabela, capaz de identificar unicamente cada registro. No entanto, uma SK não possui nenhuma relação semântica com os demais atributos de uma tabela. É apenas um valor, geralmente inteiro e sequencial, gerado com o propósito de ser a chave primária da relação.

Kimball e Ross (2010) recomendam que todas as Tabelas Dimensão tenham uma SK, mesmo aquelas com tratamento SCD do Tipo 1. Isto irá isolar o DW de surpresas ao incorporar novos dados provenientes de fontes com suas próprias ideias sobre chaves. Ainda de acordo com os autores, a utilização de SKs tornará o banco de dados mais rápido.

Existem momentos em que a utilização de SKs em Tabelas Fato (FSKs) é importante. Embora FSKs não tenham utilidade para o negócio, elas possuem diversas vantagens internamente para o DW:

- FSKs identificam um fato unicamente;
- Como FSKs são inseridas sequencialmente, uma rotina de inserção de novos registros terá as FSKs em uma faixa contínua;
- Uma FSK permite a substituição de update por insert-deletes;
- Uma FSK pode se tornar uma chave estrangeira em uma tabela fato de menor granularidade.

Uma NK pode não ser durável, visto que não teremos controle sobre possíveis mudanças de chave nas fontes de origem. NKs podem ser originárias de mais de uma fonte, e neste caso, cada fonte de dados pode utilizar uma NK diferente para um mesmo processo de negócio.

Data Profiling

Ao longo do processo de modelagem dimensional, a equipe precisa desenvolver uma compreensão cada vez maior da estrutura, conteúdo, relacionamentos e regras de derivação dos dados de origem. É necessário verificar se os dados estão em um estado utilizável, ou se ao menos suas falhas podem ser gerenciadas, entendendo o que é necessário para convertê-los para o modelo dimensional. O objetivo da atividade de *data profiling* (verificação do perfil dos dados) é a exploração do conteúdo e dos relacionamentos reais dos dados no sistema de origem, ao invés de depender de documentação, talvez incompleta ou desatualizada.