# Capítulo 1: Data Warehousing, Business Intelligence, Primer da modelagem dimensional

Este primeiro capítulo estabelece as bases para os capítulos seguintes. Nós começamos considerando sistemas de data warehousing e business intelligence (DW / BI) de uma perspectiva de alto nível. Você pode ficar desapontado ao saber que não começamos com tecnologia e ferramentas - em primeiro lugar, o sistema DW / BI deve considerar a necessidades do negócio. Com as necessidades do negócio firmemente em mãos, trabalhamos para trás através dos designs lógicos e físicos, juntamente com as decisões sobre tecnologia e ferramentas.

Nós direcionamos os investimentos para os objetivos de data warehousing e business inteligência neste capítulo, observando as estranhas semelhanças entre as responsabilidades de um gerente de DW / BI e de um editor. Com esta grande perspectiva, exploramos conceitos centrais de modelagem dimensional e estabelecer vocabulário fundamental. A partir daí este capítulo discute as principais componentes da arquitetura Kimball DW / BI, juntamente com uma comparação de abordagens arquitetônicas; felizmente, há um papel para a modelagem dimensional independentemente da sua persuasão arquitetônica. Finalmente, revisamos as dimensões mitos de modelagem. No final deste capítulo, você apreciará a necessidade para ser meio DBA (administrador de banco de dados) e meio MBA (analista de negócios) ao abordar seu projeto DW / BI.

O Capítulo 1 discute os seguintes conceitos:

* Objetivos voltados para os negócios de data warehousing e business intelligence
* Metáfora de publicação para sistemas DW / BI
* Conceitos centrais de modelagem dimensional e vocabulário, incluindo fatos tabelas e de dimensão
* Componentes e princípios da arquitetura Kimball DW / BI
* Comparação de arquiteturas alternativas DW / BI, e o papel da modelagem dimensional dentro de cada caso
* Mal-entendidos sobre modelagem dimensional

## Diferentes mundos de Captura de Dados e Análise de dados

Um dos ativos mais importantes de qualquer organização é sua informação. Este ativo é quase sempre usado para dois propósitos: manutenção de registros operacionais e análise tomando uma decisão. Simplesmente falando, os sistemas operacionais são onde você coloca os dados, e o sistema DW / BI é onde você obtém os dados.

Usuários de um sistema operacional giram as engrenagens da organização. Recebem pedidos, cadastram novos clientes, monitoram o status das atividades operacionais e registram reclamações. Os sistemas operacionais são otimizados para processar transações rapidamente. Esses sistemas quase sempre lidam com um registro de transação por vez. Eles previsivelmente realizam as mesmas tarefas operacionais várias vezes, executando os processos de negócios da organização. Dado esse foco de execução, os sistemas operacionais normalmente não mantêm o histórico, mas atualizam os dados para refletir o estado mais atual.

Os usuários de um sistema DW / BI, por outro lado, observam as rodas da organização se voltarem para avaliar o desempenho. Eles contam os novos pedidos e os comparam com os pedidos da semana anterior, perguntando por que os novos clientes se inscreveram e o que os clientes reclamaram. Eles se preocupam se os processos operacionais estão funcionando corretamente. Embora eles precisem de dados detalhados para suportar suas perguntas em constante mudança, os usuários do DW / BI quase nunca lidam com uma transação de cada vez. Esses sistemas são otimizados para consultas de alto desempenho, pois as perguntas dos usuários geralmente exigem que centenas ou centenas de milhares de transações sejam pesquisadas e compactadas em um conjunto de respostas. Para complicar ainda mais, os usuários de um sistema DW / BI normalmente exigem que o contexto histórico seja preservado para avaliar com precisão o desempenho da organização ao longo do tempo.

Na primeira edição do The Data Warehouse Toolkit (Wiley, 1996), Ralph Kimball dedicou um capítulo inteiro para descrever a dicotomia entre os mundos do processamento peracional e do data warehousing. Neste momento, é amplamente reconhecido que o sistema DW / BI tem necessidades, clientes, estruturas e ritmos profundamente diferentes do que os sistemas operacionais de registro. Infelizmente, ainda encontramos supostos sistemas DW / BI que são meras cópias dos sistemas operacionais de registro armazenados em uma plataforma de hardware separada. Embora esses ambientes possam abordar a necessidade de isolar os ambientes operacionais e analíticos por motivos de desempenho, eles não fazem nada para abordar as outras diferenças inerentes entre os dois tipos de sistemas. Os usuários de negócios não ficam impressionados com a usabilidade e o desempenho fornecidos por esses pseudo-armazéns de dados; esses impostores fazem um desserviço ao DW / BI porque não reconhecem que seus usuários têm necessidades drasticamente diferentes dos usuários do sistema operacional.

## Objetivos do Data Warehousing e Business Intelligence

Antes de nos aprofundarmos nos detalhes da modelagem dimensional, é útil focar nos objetivos fundamentais de data warehousing e business intelligence. Os objetivos podem ser facilmente desenvolvidos, percorrendo os corredores de qualquer organização e ouvindo a gestão de negócios. Esses temas recorrentes existem há mais de três décadas:

* "Nós coletamos toneladas de dados, mas não podemos acessá-los."
* "Precisamos cortar e dividir os dados de todas as maneiras."
* "As pessoas de negócios precisam obter facilmente os dados".
* "Apenas me mostre o que é importante."
* "Passamos reuniões inteiras discutindo quem tem os números certos em vez de tomar decisões."
* "Queremos que as pessoas usem informações para apoiar tomadas de decisão mais baseadas em fatos."

Com base em nossa experiência, essas preocupações ainda são tão universais que conduzem os requisitos básicos para o sistema DW / BI. Agora, transforme essas cotações de gerenciamento de negócios em requisitos.

* **O sistema DW / BI deve tornar as informações facilmente acessíveis.** O conteúdo do sistema DW / BI deve ser compreensível. Os dados devem ser intuitivos e óbvios para o usuário corporativo, não apenas para o desenvolvedor. As estruturas e os rótulos dos dados devem imitar os processos de pensamento e o vocabulário dos usuários de negócios. Os usuários de negócios desejam separar e combinar dados analíticos em combinações infinitas. As ferramentas e aplicativos de business intelligence que acessam os dados devem ser simples e fáceis de usar. Eles também devem retornar os resultados da consulta ao usuário com tempos de espera mínimos. Podemos resumir esse requisito simplesmente dizendo simples e rápido.
* **O sistema DW / BI deve apresentar informações de forma consistente.** Os dados no sistema DW / BI devem ser confiáveis. Os dados devem ser cuidadosamente montados a partir de uma variedade de fontes, limpos, com garantia de qualidade e liberados somente quando estiverem aptos para o consumo do usuário. A consistência também implica que rótulos e definições comuns para o conteúdo do sistema DW / BI são usados em fontes de dados. Se duas medidas de desempenho tiverem o mesmo nome, elas devem significar a mesma coisa. Por outro lado, se duas medidas não significam a mesma coisa, elas devem ser rotuladas de forma diferente.
* **O sistema DW / BI deve se adaptar à mudança.** As necessidades do usuário, as condições comerciais, os dados e a tecnologia estão sujeitos a alterações. O sistema DW / BI deve ser projetado para lidar com essa inevitável alteração de forma a não invalidar dados ou aplicativos existentes. Os dados e aplicativos existentes não devem ser alterados ou interrompidos quando a comunidade de negócios fizer novas perguntas ou novos dados forem adicionados ao warehouse. Finalmente, se os dados descritivos do sistema DW / BI precisarem ser modificados, você deve contabilizar apropriadamente as alterações e tornar essas alterações transparentes para os usuários.
* **O sistema DW / BI deve apresentar informações em tempo hábil.** Como o sistema DW / BI é usado mais intensamente para decisões operacionais, os dados brutos podem precisar ser convertidos em informações acionáveis em horas, minutos ou até segundos. A equipe de DW / BI e os usuários de negócios precisam ter expectativas realistas sobre o que significa entregar dados quando há pouco tempo para limpá-los ou validá-los.
* **O sistema DW / BI deve ser um bastião seguro que proteja os ativos de informação.** As joias da coroa informativas de uma organização são armazenadas no data warehouse. No mínimo, o warehouse provavelmente contém informações sobre o que você está vendendo e a que preço - detalhes potencialmente prejudiciais nas mãos das pessoas erradas. O sistema DW / BI deve controlar efetivamente o acesso às informações confidenciais da organização.
* **O sistema DW / BI deve servir como base autoritativa e confiável para melhorar a tomada de decisões.** O data warehouse deve ter os dados corretos para dar suporte à tomada de decisões. Os resultados mais importantes de um sistema DW / BI são as decisões tomadas com base nas evidências analíticas apresentadas; essas decisões fornecem o impacto e o valor do negócio atribuíveis ao sistema DW / BI. O rótulo original que antecede o DW / BI ainda é a melhor descrição do que você está projetando: um sistema de suporte à decisão.
* **A comunidade empresarial deve aceitar o sistema DW / BI para considerá-lo bem-sucedido.** Não importa que você tenha criado uma solução elegante usando os melhores produtos e plataformas. Se a comunidade de negócios não abraçar o ambiente DW / BI e usá-lo ativamente, você falhou no teste de aceitação. Ao contrário de uma implementação de sistema operacional em que os usuários corporativos não têm outra opção a não ser usar o novo sistema, o uso de DW / BI às vezes é opcional. Os usuários corporativos adotarão o sistema DW / BI se ele for a fonte “simples e rápida” de informações acionáveis.

Embora cada requisito nesta lista seja importante, os dois últimos são os mais críticos e, infelizmente, frequentemente os mais negligenciados. Armazenamento de dados e business intelligence bem-sucedidos exigem mais do que um arquiteto, técnico, modelador ou administrador de banco de dados estelar. Com uma iniciativa de DW / BI, você tem um pé em sua zona de conforto de tecnologia da informação (TI), enquanto o outro pé está na região desconhecida dos usuários corporativos. Você deve ficar entre os dois, modificando algumas habilidades testadas e verdadeiras para se adaptar às demandas exclusivas do DW / BI. Claramente, você precisa trazer um espectro de habilidades para a festa se comportar como se você fosse um DBA / MBA híbrido.

### Publicação de Metáfora para Gerentes de DW / BI

Com os objetivos do DW / BI como pano de fundo, vamos comparar as responsabilidades dos gerentes de DW / BI com as de um editor-chefe de publicação. Como editor de uma revista de alta qualidade, você teria ampla liberdade para gerenciar o conteúdo, o estilo e a entrega da revista. Qualquer pessoa com esse cargo provavelmente abordaria as seguintes atividades:

* Entenda os leitores:
  + Identifique suas características demográficas.
  + Descubra o que os leitores querem neste tipo de revista.
  + Identifique os "melhores" leitores que renovarão suas assinaturas e comprarão produtos dos anunciantes da revista.
  + Encontre potenciais novos leitores e torne-os cientes da revista.
* Garanta que a revista agrade aos leitores:
  + Escolha conteúdo de revista interessante e atraente.
  + Tome decisões de layout e renderização que maximizem o prazer dos leitores.
  + Defenda padrões de escrita e edição de alta qualidade enquanto adota um estilo de apresentação consistente.
  + Monitore continuamente a precisão dos artigos e das reivindicações dos anunciantes.
  + Adapte-se à mudança de perfis de leitores e à disponibilidade de novas entradas de uma rede de escritores e colaboradores.
* Sustentar a publicação:
  + Atraia anunciantes e gerencie a revista de maneira lucrativa.
  + Publique a revista regularmente.
  + Mantenha a confiança dos leitores.
  + Mantenha os donos de empresas felizes.

Você também pode identificar itens que não deveriam ser objetivos do editor-chefe da revista, como construir a revista em torno de uma tecnologia de impressão específica ou colocar exclusivamente a energia da administração em eficiência operacional, como impor um estilo de redação técnica que os leitores não tenham. Facilmente entender ou criar um layout complexo e intrincado que seja difícil de ler.

Ao construir o negócio de publicações com base em servir os leitores efetivamente, a revista provavelmente terá sucesso. Por outro lado, percorra a lista e imagine o que acontece se você omitir qualquer item único; em última análise, a revista teria sérios problemas.

Existem fortes paralelos que podem ser traçados entre ser um editor convencional e ser um gerente de DW / BI. Impulsionados pelas necessidades do negócio, os gerentes do DW / BI devem publicar dados que foram coletados de uma variedade de fontes e editados para qualidade e consistência. A principal responsabilidade é atender os leitores, também conhecidos como usuários corporativos. A metáfora da publicação ressalta a necessidade de focar seus clientes em vez de meramente focar nos produtos e processos. Embora você use a tecnologia para fornecer o sistema DW / BI, a tecnologia é, na melhor das hipóteses, um meio para um fim. Como tal, a tecnologia e as técnicas usadas para construir o sistema não devem aparecer diretamente em suas responsabilidades de trabalho.

Agora reformule as responsabilidades do editor de revista como gerente de DW / BI responsabilidades:

* Entenda os usuários de negócios:
  + Entenda suas responsabilidades de trabalho, metas e objetivos.
  + Determine as decisões que os usuários corporativos desejam tomar com a ajuda do sistema DW / BI.
  + Identifique os "melhores" usuários que tomam decisões eficazes e de alto impacto.
  + Encontre novos usuários em potencial e conscientize-os sobre os recursos do sistema DW / BI.
* Forneça informações e análises de alta qualidade, relevantes e acessíveis para os usuários de negócios:
  + Escolha os dados mais robustos e acionáveis para apresentar no sistema DW / BI, cuidadosamente selecionados do vasto universo de possíveis fontes de dados em sua organização.
  + Torne as interfaces e os aplicativos do usuário simples e orientados a modelos, correspondendo explicitamente aos perfis de processamento cognitivo dos usuários.
  + Verifique se os dados são precisos e confiáveis, rotulando-os de maneira consistente em toda a empresa.
  + Monitore continuamente a precisão dos dados e análises.
  + Adapte-se à mudança de perfis de usuários, requisitos e prioridades de negócios, juntamente com a disponibilidade de novas fontes de dados.
* Sustente o ambiente DW / BI:
  + Pegue uma parte do crédito para as decisões de negócios feitas usando o sistema DW / BI, e use esses sucessos para justificar as despesas com pessoal e em curso.
  + Atualize o sistema DW / BI regularmente.
  + Mantenha a confiança dos usuários corporativos.
  + Mantenha os usuários de negócios, os patrocinadores executivos e o gerenciamento de TI satisfeitos.

Se você fizer um bom trabalho com todas essas responsabilidades, será um ótimo gerente de DW / BI! Por outro lado, percorra a lista e imagine o que acontece se você omitir qualquer item individual. Em última análise, o ambiente teria sérios problemas. Agora, compare essa visão do trabalho de um gerente de DW / BI com a descrição de seu próprio trabalho. As chances são de que a lista anterior é mais orientada para problemas de usuários e negócios e pode nem soar como um trabalho na área de TI. Em nossa opinião, isso é o que torna interessante o armazenamento de dados e a inteligência de negócios.

## Introdução à Modelagem Dimensional

Agora que você entende os objetivos do sistema DW / BI, vamos considerar os fundamentos da modelagem dimensional. A modelagem dimensional é amplamente aceita como a técnica preferida para a apresentação de dados analíticos, pois atende a dois requisitos simultâneos:

* Forneça dados que sejam compreensíveis para os usuários corporativos.
* Entregue desempenho de consulta rápida.

A modelagem dimensional é uma técnica de longa data para tornar os bancos de dados simples. Em caso após caso, por mais de cinco décadas, as organizações de TI, consultores e usuários de negócios gravitaram naturalmente para uma estrutura dimensional simples para atender à necessidade humana fundamental de simplicidade. A simplicidade é fundamental porque garante que os usuários possam entender facilmente os dados, além de permitir que o software navegue e forneça resultados de maneira rápida e eficiente.

Imagine um executivo que descreve seu negócio como: "Vendemos produtos em vários mercados e medimos nosso desempenho ao longo do tempo". Os designers dimensionais ouvem atentamente a ênfase no produto, no mercado e no tempo. A maioria das pessoas acha intuitivo pensar em um negócio como um cubo de dados, com as bordas rotuladas como produto, mercado e tempo. Imagine cortar e cortar cada uma dessas dimensões. Pontos dentro do cubo são onde as medidas, como volume de vendas ou lucro, para essa combinação de produto, mercado e tempo são armazenadas. A capacidade de visualizar algo tão abstrato quanto um conjunto de dados de maneira concreta e tangível é o segredo da compreensibilidade. Se essa perspectiva parece simples demais, bom! Um modelo de dados que começa simples tem a chance de permanecer simples no final do design. Um modelo que começa complicado certamente será extremamente complicado no final, resultando em lentidão no desempenho das consultas e na rejeição dos usuários corporativos. Albert Einstein capturou a filosofia básica que guia o design dimensional quando disse: "Torne tudo o mais simples possível, mas não mais simples".

Embora modelos dimensionais sejam frequentemente instanciados em sistemas de gerenciamento de banco de dados relacional, eles são bastante diferentes dos modelos de terceira forma normal (3NF) que procuram remover redundâncias de dados. Estruturas 3NF normalizadas dividem os dados em muitas entidades distintas, cada uma das quais se torna uma tabela relacional. Um banco de dados de pedidos de venda pode começar com um registro para cada linha de pedido, mas se transformar em um diagrama de teia de aranha complexo como um modelo 3NF, talvez consistindo em centenas de tabelas normalizadas.

O setor às vezes se refere a modelos 3NF como modelos de relação de entidade (ER). Diagramas de relação de entidade (diagramas ER ou ERDs) são desenhos que comunicam as relações entre tabelas. Modelos 3NF e dimensionais podem ser representados em ERDs porque ambos consistem em tabelas relacionais unidas; a principal diferença entre 3NF e modelos dimensionais é o grau de normalização. Como ambos os tipos de modelo podem ser apresentados como ERDs, nos abstivemos de nos referir a modelos 3NF como modelos ER; em vez disso, os chamamos de modelos normalizados para minimizar a confusão.

Estruturas 3NF normalizadas são imensamente úteis no processamento operacional porque uma transação de atualização ou inserção toca o banco de dados em apenas um local. Modelos normalizados, no entanto, são muito complicados para consultas de BI. Os usuários não conseguem entender, navegar ou lembrar de modelos normalizados que se assemelham a um mapa do sistema de vias expressas de Los Angeles. Da mesma forma, a maioria dos sistemas de gerenciamento de banco de dados relacional não consegue consultar com eficiência um modelo normalizado; a complexidade das consultas imprevisíveis dos usuários sobrecarrega os otimizadores de banco de dados, resultando em um desempenho de consulta desastroso. O uso de modelagem normalizada na área de apresentação do DW / BI vence a recuperação intuitiva e de alto desempenho dos dados. Felizmente, a modelagem dimensional aborda o problema de esquemas excessivamente complexos na área de apresentação.

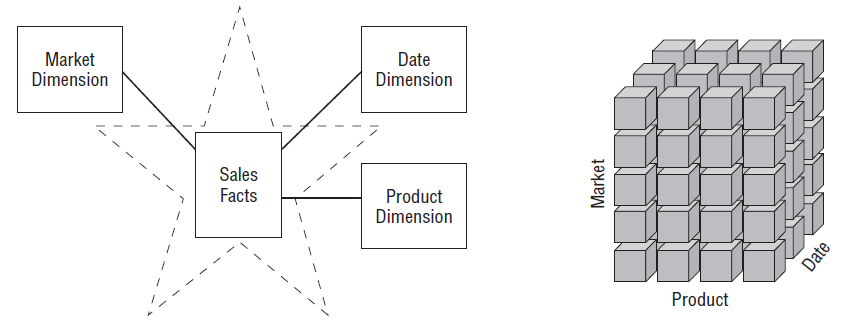
|  |
| --- |
| NOTA |
| Um modelo dimensional contém as mesmas informações que um modelo normalizado, mas compacta os dados em um formato que ofereça capacidade de compreensão do usuário, desempenho de consulta e resiliência para alterar. |

### Esquemas em estrela versus cubos OLAP

Os modelos dimensionais implementados em sistemas de gerenciamento de banco de dados relacionais são chamados de esquemas em estrela devido à sua semelhança com uma estrutura em forma de estrela. Os modelos dimensionais implementados em ambientes de bancos de dados multidimensionais são chamados de cubos de processamento analítico online (OLAP), conforme ilustrado na Figura 1-1.

Se seu ambiente DW / BI incluir esquemas em estrela ou cubos OLAP, ele aproveitará os conceitos dimensionais. Ambas as estrelas e cubos têm um design lógico comum com dimensões reconhecíveis; no entanto, os diferenciais de implementação física.

Quando os dados são carregados em um cubo OLAP, eles são armazenados e indexados usando formatos e técnicas projetados para dados dimensionais. As agregações de desempenho ou tabelas de resumo pré-calculadas são frequentemente criadas e gerenciadas pelo mecanismo de cubo OLAP. Conseqüentemente, os cubos oferecem um desempenho de consulta superior devido aos cálculos prévios, estratégias de indexação e outras otimizações. Os usuários de negócios podem detalhar ou adicionar adicionando ou removendo atributos de suas análises com excelente desempenho sem emitir novas consultas. Os cubos OLAP também fornecem funções analíticas mais robustas que excedem as disponíveis com o SQL. A desvantagem é que você paga um preço de desempenho de carga para esses recursos, especialmente com grandes conjuntos de dados.



**Figura 1-1Esquema Estrela versus cubo OLAP**

Felizmente, a maioria das recomendações contidas neste livro pertence independentemente da plataforma de banco de dados relacional versus multidimensional. Embora os recursos da tecnologia OLAP estejam melhorando continuamente, geralmente recomendamos que informações atômicas detalhadas sejam carregadas em um esquema em estrela; Os cubos OLAP opcionais são então preenchidos do esquema em estrela. Por essa razão, a maioria das técnicas de modelagem dimensional neste livro é expressa em termos de um esquema estrela relacional.

### Considerações de implantação OLAP

Aqui estão algumas coisas para manter em mente se você implantar dados em cubos OLAP:

* Um esquema em estrela hospedado em um banco de dados relacional é uma boa base física para a criação de um cubo OLAP e é geralmente considerado uma base mais estável para backup e recuperação.
* Tradicionalmente, os cubos OLAP são conhecidos por vantagens extremas de desempenho em relação aos RDBMSs, mas essa distinção se tornou menos importante com os avanços no hardware, como aplicativos e bancos de dados em memória, e software RDBMS, como bancos de dados colunares.
* Estruturas de dados de cubos OLAP são mais variáveis entre fornecedores diferentes que DBMSs relacionais, portanto, os detalhes da implantação final geralmente dependem de qual fornecedor OLAP é escolhido. Normalmente, é mais difícil portar aplicativos de BI entre diferentes ferramentas OLAP do que a porta de aplicativos de BI em diferentes bancos de dados relacionais.
* Os cubos OLAP geralmente oferecem opções de segurança mais sofisticadas do que os RDBMSs, como limitar o acesso a dados detalhados, mas fornecer mais acesso aberto a dados resumidos.
* O OLAP não oferece recursos de análise significativamente mais ricos do que os RDBMSs, que são sobrecarregados pelas restrições do SQL. Esta pode ser a principal justificativa para usar um produto OLAP.
* Os cubos OLAP suportam as alterações de tipo de dimensão 2 de alteração lenta (discutidas no Capítulo 5: Procurement), mas os cubos geralmente precisam ser reprocessados parcial ou totalmente sempre que os dados forem sobregravados usando técnicas de dimensão alternativas que mudam lentamente.
* Os cubos OLAP suportam graciosamente tabelas de fatos de instantâneos de transações e periódicos, mas não manipulam tabelas de fatos de instantâneo acumuladas devido às limitações de sobrescrever dados descritos no ponto anterior.
* Os cubos OLAP geralmente suportam hierarquias irregulares complexas de profundidade indeterminada, como organogramas ou listas de materiais, usando uma sintaxe de consulta nativa superior às abordagens necessárias para os RDBMSs.
* Os cubos OLAP podem impor restrições detalhadas na estrutura das chaves de dimensão que implementam hierarquias de detalhamento comparadas aos bancos de dados relacionais.
* Alguns produtos OLAP não permitem funções dimensionais ou aliases, exigindo, portanto, que dimensões físicas separadas sejam definidas.

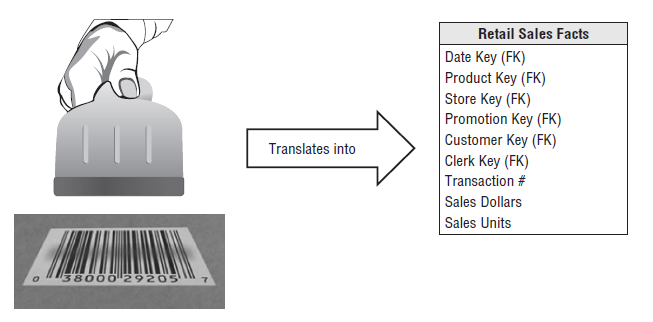
Voltaremos ao mundo da modelagem dimensional em uma plataforma relacional à medida que consideramos os dois principais componentes de um esquema em estrela.

### Tabelas de fatos para medições

A tabela de fatos em um modelo dimensional armazena as medidas de desempenho resultantes dos eventos do processo de negócios de uma organização. Você deve se esforçar para armazenar os dados de medição de baixo nível resultantes de um processo de negócios em um modelo unidimensional. Como os dados de medição são, em sua maioria, o maior conjunto de dados, eles não devem ser replicados em vários locais para várias funções organizacionais em toda a empresa. Permitir que usuários corporativos de várias organizações acessem um único repositório centralizado para cada conjunto de dados de medição garante o uso de dados consistentes em toda a empresa.

O termo fato representa uma medida de negócios. Imagine estar no mercado observando os produtos sendo vendidos e anotando a quantidade de unidades e o valor das vendas em dólar de cada produto em cada transação de vendas. Essas medidas são capturadas conforme os produtos são digitalizados no registro, conforme ilustrado na Figura 2.

Cada linha em uma tabela de fatos corresponde a um evento de medição. Os dados em cada linha estão em um nível de detalhe específico, chamado de granularidade, como uma linha por produto vendido em uma transação de venda. Um dos principais princípios da modelagem dimensional é que todas as linhas de medidas em uma tabela de fatos devem estar no mesmo nível. Ter a disciplina para criar tabelas de fatos com um único nível de detalhe garante que as medições não sejam duplamente contadas.



**Figura 2 - Os eventos de medição do processo de negócios são convertidos em tabelas de fatos.**

|  |
| --- |
| Nota |
| 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. Tudo o resto se baseia nesta fundação. |

Os fatos mais úteis são numéricos e aditivos, como o valor das vendas em dólar. Ao longo deste livro, usaremos dólares como moeda padrão para tornar os exemplos de estudo de caso mais tangíveis. Você pode substituir sua própria moeda local, se não for em dólares.

A aditividade é crucial porque os aplicativos de BI raramente recuperam uma única linha da tabela de fatos. Em vez disso, eles trazem de volta centenas, milhares ou até milhões de linhas reais por vez, e a coisa mais útil para fazer com tantas linhas é adicioná-las. Não importa como o usuário divide os dados na Figura 1-2, as unidades de vendas e os dólares somam um total válido. Você verá que os fatos às vezes são semi-aditivos ou mesmo não-aditivos. Fatos semi-aditivos, como saldos de contas, não podem ser somados na dimensão de tempo. Fatos não aditivos, como preços unitários, nunca podem ser adicionados. Você é forçado a usar contagens e médias ou é reduzido a imprimir as linhas de fatos uma de cada vez - um exercício impraticável com uma tabela de fatos de um bilhão de linhas.

Os fatos são frequentemente descritos como continuamente valorizados para ajudar a classificar o que é um fato em relação a um atributo de dimensão. O fato da quantia de vendas em dólares é continuamente avaliado neste exemplo, porque pode assumir virtualmente qualquer valor dentro de um amplo intervalo. Como observador, você deve se destacar no mercado e esperar pela medição antes de ter qualquer ideia de qual será o valor.

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 projetista deve fazer todo o possível para colocar os dados textuais em dimensões, onde eles possam ser correlacionados mais efetivamente com os outros atributos da dimensão textual e consumir muito menos espaço. 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 dimensões. 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 quase impossível analisá-lo.

Consultando a tabela de fatos da amostra na Figura 1-2, se não houver atividade de vendas para um determinado produto, você não colocará nenhuma linha na tabela. É importante que você não tente preencher a tabela de fatos com zeros representando nenhuma atividade, porque esses zeros sobrecarregariam a maioria das tabelas de fatos. Ao incluir apenas a atividade verdadeira, as tabelas de fatos tendem a ser bastante esparsas. Apesar de sua escassez, as tabelas de fatos geralmente representam 90% ou mais do espaço total consumido por um modelo dimensional. As tabelas de fatos tendem a ser profundas em termos do número de linhas, mas estreitam em termos do número de colunas. Dado o seu tamanho, você deve ser criterioso sobre a utilização do espaço de tabela de fatos.

Como exemplos são desenvolvidos ao longo deste livro, você verá que todos os grãos da tabela de fatos se enquadram em uma das três categorias: transação, instantâneo periódico e acumular instantâneo. Tabelas de fatos de grãos de transações são as mais comuns. Introduziremos as tabelas de fatos da transação no Capítulo 3: Vendas no varejo e capturas instantâneas periódicas e acumuladas no Capítulo 4: Inventário.

Todas as tabelas de fatos possuem duas ou mais chaves estrangeiras (consulte a notação FK na Figura 2) que se conectam às chaves primárias das tabelas de dimensão. Por exemplo, a chave do produto na tabela de fatos sempre corresponde a uma chave de produto específica na tabela de dimensões do produto. Quando todas as chaves na tabela de fatos correspondem corretamente às respectivas chaves primárias nas tabelas de dimensão correspondentes, as tabelas satisfazem a integridade referencial. Você acessa a tabela de fatos por meio das tabelas de dimensões associadas a ela.

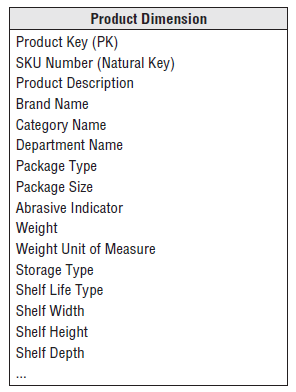
A tabela de fatos geralmente tem sua própria chave primária composta de um subconjunto das chaves estrangeiras. Essa chave é geralmente chamada de chave composta. Cada tabela que possui uma chave composta é uma tabela de fatos. Tabelas de fatos expressam relacionamentos muitos-para-muitos. Todos os outros são tabelas de dimensões.

Geralmente, há algumas dimensões que juntas identificam exclusivamente cada linha da tabela de fatos. Depois que esse subconjunto da lista de dimensões geral for identificado, o resto das dimensões assumirá um único valor no contexto da chave primária da linha da tabela de fatos. Em outras palavras, eles acompanham o passeio.

### Tabelas de dimensão para contexto descritivo

As tabelas de dimensão são complementares integrais a uma tabela de fatos. As tabelas de dimensão contêm o contexto textual associado a um evento de medição do processo de negócios. Eles descrevem o “quem, o que, onde, quando, como e por que” associado ao evento.

Conforme ilustrado na Figura 3, as tabelas de dimensões geralmente têm muitas colunas ou atributos. Não é incomum que uma tabela de dimensões tenha 50 a 100 atributos; embora, algumas tabelas de dimensões tenham naturalmente apenas alguns atributos. As tabelas de dimensões tendem a ter menos linhas que as tabelas de fatos, mas podem ser largas com muitas colunas de texto grandes. Cada dimensão é definida por uma única chave primária (consulte a notação PK na Figura 3), que serve como base para a integridade referencial com qualquer tabela de fatos para a qual ela é unida.



**Figura 3- As tabelas de dimensões contêm características descritivas dos substantivos do processo de negócios.**

Atributos de dimensão servem como fonte primária de restrições de consulta, agrupamentos e rótulos de relatório. Em uma solicitação de consulta ou relatório, os atributos são identificados como por palavras. Por exemplo, quando um usuário deseja ver as vendas em dólares por marca, a marca deve estar disponível como um atributo de dimensão.

Os atributos da tabela de dimensões desempenham um papel vital no sistema DW / BI. Como eles são a fonte de praticamente todas as restrições e rótulos de relatórios, os atributos de dimensão são essenciais para tornar o sistema DW / BI utilizável e compreensível. Atributos devem consistir em palavras reais, em vez de abreviações enigmáticas. Você deve se esforçar para minimizar o uso de códigos em tabelas de dimensões, substituindo-os por atributos textuais mais detalhados.Você já pode ter treinado os usuários empresariais para memorizar códigos operacionais, mas, no futuro, minimizar sua dependência de notas em miniatura conectadas ao monitor por código traduções. Você deve criar decodificações padrão para os códigos operacionais disponíveis como atributos de dimensão para fornecer rotulação consistente em consultas, relatórios e aplicativos de BI. Os valores de decodificação nunca devem ser enterrados nos aplicativos de relatório, onde a inconsistência é inevitável.

Às vezes, códigos operacionais ou identificadores têm um significado comercial legítimo para os usuários ou são obrigados a se comunicar de volta ao mundo operacional. Nesses casos, os códigos devem aparecer como atributos de dimensão explícita, além dos descritores de texto correspondentes e fáceis de usar. Os códigos operacionais às vezes possuem inteligência embutida neles. Por exemplo, os dois primeiros 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 substrings dentro dos códigos operacionais, retire os significados incorporados e apresente-os aos usuários como atributos de dimensão separados que podem ser facilmente filtrados, agrupados ou relatados.

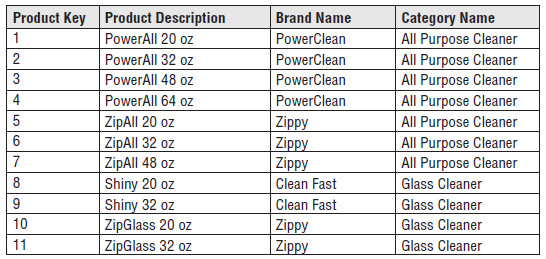
De muitas maneiras, o data warehouse é tão bom quanto os atributos de dimensão; o poder analítico do ambiente DW / BI é diretamente proporcional à qualidade e profundidade dos atributos de dimensão. Quanto mais tempo gasto fornecendo atributos com terminologia de negócios detalhada, melhor. Quanto mais tempo gasto preenchendo os valores de domínio em uma coluna de atributo, melhor. Quanto mais tempo gasto para garantir a qualidade dos valores em uma coluna de atributos, melhor. Atributos de dimensão robustos fornecem recursos analíticos de slicing-and-dicing.

|  |
| --- |
| Nota |
| As dimensões fornecem os pontos de entrada para os dados e os rótulos e agrupamentos finais em todas as análises de DW / BI. |

Ao triar os dados de origem operacionais, às vezes não está claro se um elemento de dados numéricos é um atributo de fato ou de dimensão. Você frequentemente toma a decisão perguntando se a coluna é uma medida que assume muitos valores e participa de cálculos (tornando-se um fato) ou é uma descrição discretamente valorizada que é mais ou menos constante e participa de restrições e rótulos de linhas (tornando é um atributo dimensional). Por exemplo, o custo padrão de um produto parece ser um atributo constante do produto, mas pode ser alterado com tanta frequência que você decide que se trata mais de um fato medido. Ocasionalmente, você não pode ter certeza da classificação; É possível modelar o elemento de dados de qualquer forma (ou ambas as formas) como uma questão da prerrogativa do designer.

|  |
| --- |
| Nota |
| O dilema do designer de saber se uma quantidade numérica é um fato ou um atributo de dimensão raramente é uma decisão difícil. Observações numéricas de valor contínuo são quase sempre fatos; Observações numéricas discretas extraídas de uma lista pequena são quase sempre atributos de dimensão. |

A Figura 1-4 mostra que as tabelas de dimensões geralmente representam relacionamentos hierárquicos. Por exemplo, os produtos se acumulam em marcas e depois em categorias. Para cada linha na dimensão do produto, você deve armazenar a descrição de marca e categoria associada. A informação descritiva hierárquica é armazenada de forma redundante no espírito de facilidade de uso e desempenho da consulta. Você deve resistir ao desejo talvez habitual de normalizar os dados, armazenando apenas o código da marca na dimensão do produto e criando uma tabela de pesquisa de marca separada, e também para a descrição da categoria em uma tabela de pesquisa de categoria separada. Essa normalização é chamada de snowfl aking. Em vez do terceiro formulário normal, as tabelas de dimensões normalmente são altamente desnormalizadas com relacionamentos muitos-para-um em uma única tabela de dimensões. Como as tabelas de dimensões normalmente são geometricamente menores que as tabelas de fatos, melhorar a eficiência de armazenamento por meio da normalização ou do snowfl aking praticamente não tem impacto no tamanho geral do banco de dados. Você deve quase sempre trocar o espaço da tabela de dimensões pela simplicidade e acessibilidade.

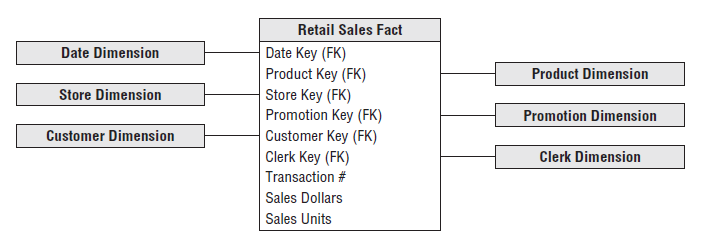


**Figura 4 - Amostra de linhas de uma tabela de dimensões com hierarquias desordenadas.**

Ao contrário do folclore popular, Ralph Kimball não inventou os termos fato e dimensão. Tanto quanto se pode determinar, a dimensão e a terminologia dos fatos se originaram de um projeto conjunto de pesquisa conduzido pela General Mills e pela Universidade de Dartmouth na década de 1960. Na década de 1970, AC Nielsen e IRI usaram os termos consistentemente para descrever seus dados distribuídos por sindicatos e gravitaram para modelos dimensionais para simplificar a apresentação de suas informações analíticas. Eles entenderam que seus dados não seriam usados, a menos que fossem empacotados de maneira simples. Provavelmente, é preciso dizer que nenhuma pessoa inventou a abordagem dimensional. É uma força irresistível na criação de bancos de dados que sempre resulta quando o designer coloca a compreensibilidade e o desempenho como as metas mais altas.

### Fatos e Dimensões cadastradas em um esquema em estrela

Agora que você entende as tabelas de fatos e dimensões, é hora de reunir os blocos de construção em um modelo dimensional, conforme mostrado na Figura 1-5. Cada processo de negócios é representado por um modelo dimensional que consiste em uma tabela de fatos contendo as medidas numéricas do evento envolvidas por um halo de tabelas de dimensões que contêm o contexto textual que era verdadeiro no momento em que o evento ocorreu. Essa estrutura estrela-característica é freqüentemente chamada de junção estelar, um termo que remonta aos primeiros dias dos bancos de dados relacionais.



**Figura 5 - Tabelas de fatos e dimensões em um modelo dimensional.**

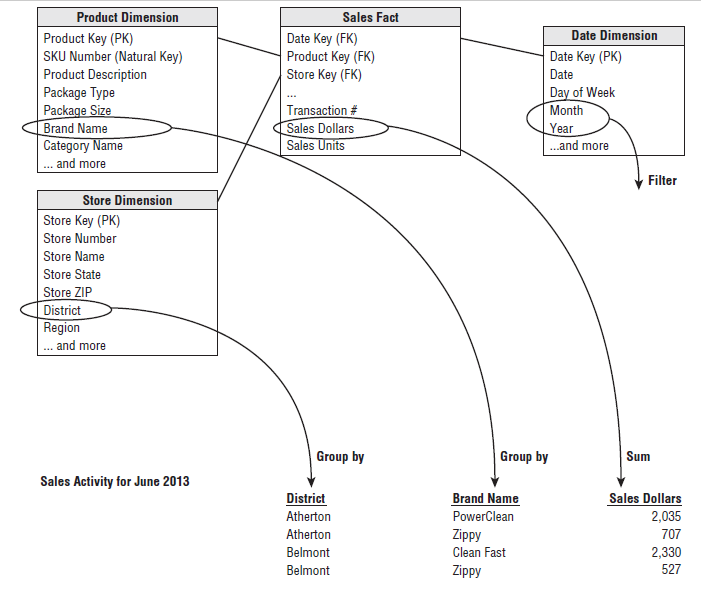
A primeira coisa a notar sobre o esquema dimensional é sua simplicidade e simetria. Obviamente, os usuários de negócios se beneficiam da simplicidade porque os dados são mais fáceis de entender e navegar. O charme do design na Figura 1-5 é que ele é altamente reconhecível pelos usuários corporativos. Observamos literalmente centenas de instâncias em que os usuários concordam imediatamente que o modelo dimensional é seu negócio. Além disso, o número reduzido de tabelas e o uso de descritores de negócios significativos facilitam a navegação e reduzem a probabilidade de erros.

A simplicidade de um modelo dimensional também tem benefícios de desempenho. Os otimizadores de banco de dados processam esses esquemas simples com menos junções de maneira mais eficiente. Um mecanismo de banco de dados pode fazer fortes pressuposições sobre restringir primeiro as tabelas de dimensão altamente indexadas e, em seguida, atacar a tabela de fatos de uma só vez com o produto cartesiano das chaves da tabela de dimensão satisfazendo as restrições do usuário. Surpreendentemente, usando essa abordagem, o otimizador pode avaliar junções arbitrárias de n-way a uma tabela de fatos em uma única passagem pelo índice da tabela de fatos.

Finalmente, modelos dimensionais são graciosamente extensíveis para acomodar mudanças. A estrutura previsível de um modelo dimensional suporta mudanças inesperadas no comportamento do usuário. Cada dimensão é equivalente; Todas as dimensões são pontos de entrada simetricamente iguais na tabela de fatos. O modelo dimensional não possui preconceito interno em relação aos padrões de consulta esperados. Não há preferências para as perguntas de negócios feitas este mês em comparação com as perguntas feitas no próximo mês. Você certamente não deseja ajustar esquemas se os usuários de negócios sugerirem novas formas de analisar seus negócios.

Este livro ilustra repetidamente que os dados mais granulares ou atômicos têm amais dimensionalidade. Dados atômicos que não foram agregados são os dados mais expressivos; esses dados atômicos devem ser a base de todo design de tabela de fatos para resistir a ataques ad hoc de usuários corporativos, nos quais eles colocam consultas inesperadas. Com os modelos dimensionais, você pode adicionar dimensões completamente novas ao esquema, desde que um único valor dessa dimensão seja definido para cada linha de fato existente. Da mesma forma, você pode adicionar novos fatos à tabela de fatos, supondo que o nível de detalhe seja consistente com a tabela de fatos existente. Você pode complementar as tabelas de dimensões preexistentes com atributos novos e imprevistos. Em cada caso, tabelas existentes podem ser alteradas no local simplesmente adicionando novas linhas de dados na tabela ou executando um comando SQL ALTER TABLE. Os dados não precisariam ser recarregados e os aplicativos de BI existentes continuariam a ser executados sem produzir resultados diferentes. Examinamos essa extensibilidade graciosa dos modelos dimensionais de forma mais completa no Capítulo 3.

Outra maneira de pensar sobre a natureza complementar das tabelas de fatos e dimensões é vê-las traduzidas em um relatório. Conforme ilustrado na Figura 1-6, os atributos de dimensão fornecem os filtros e a legenda do relatório, enquanto as tabelas de fatos fornecem os valores numéricos do relatório.



**Figura 6 - Atributos dimensionais e fatos formam um relatório simples.**

Você pode visualizar facilmente o SQL que é escrito (ou mais provavelmente gerado por uma ferramenta de BI) para criar este relatório:

SELECT

store.district\_name,

product.brand,

sum(sales\_facts.sales\_dollars) AS "Sales Dollars"

FROM

store,

product,

date,

sales\_facts

WHERE

date.month\_name="January" AND

date.year=2013 AND

store.store\_key = sales\_facts.store\_key AND

product.product\_key = sales\_facts.product\_key AND

date.date\_key = sales\_facts.date\_key

GROUP BY

store.district\_name,

product.brand

Se você estudar este snippet de código linha por linha, as duas primeiras linhas da instrução SELECT identificarão os atributos de dimensão no relatório, seguidos pela métrica agregada da tabela de fatos. A cláusula FROM identifica todas as tabelas envolvidas na consulta. As duas primeiras linhas na cláusula WHERE declaram o fi ltro do relatório e o restante declara as uniões entre as tabelas de dimensão e de fatos. Finalmente, a cláusula GROUP BY estabelece a agregação no relatório.

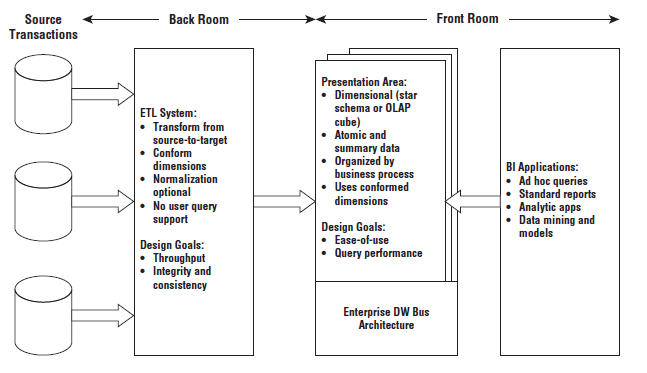
## Arquitetura DW / BI de Kimball

Vamos nos basear em sua compreensão dos sistemas DW / BI e dos fundamentos de modelagem dimensional, investigando os componentes de um ambiente DW / BI com base na arquitetura Kimball. Você precisa aprender o significado estratégico de cada componente para evitar confundir seu papel e função.

Conforme ilustrado na Figura 1-7, há quatro componentes separados e distintos a serem considerados no ambiente DW / BI: sistemas de origem operacional, sistema ETL, área de apresentação de dados e aplicativos de business intelligence.

### Sistemas de fonte operacional

Estes são os sistemas operacionais de registro que capturam as transações da empresa. Pense nos sistemas de origem como fora do data warehouse porque presumivelmente você tem pouco ou nenhum controle sobre o conteúdo e o formato dos dados nesses sistemas operacionais. As principais prioridades dos sistemas de origem são processamento de desempenho e disponibilidade. Consultas operacionais contra sistemas de origem são consultas restritas, de um registro por vez, que fazem parte do fluxo de transações normal e são severamente restritas em suas demandas no sistema operacional. É seguro assumir que os sistemas de origem não são consultados das maneiras mais abrangentes e inesperadas que os sistemas DW / BI normalmente são consultados. Os sistemas de origem mantêm poucos dados históricos; um bom data warehouse pode aliviar os sistemas de origem de grande parte da responsabilidade de representar o passado. Em muitos casos, os sistemas de origem são aplicativos para fins especiais sem qualquer compromisso de compartilhar dados comuns, como produto, cliente, geografia ou calendário, com outros sistemas operacionais da organização. Naturalmente, um sistema ERP (enterprise resource planning) de aplicação cruzada amplamente adotado ou um sistema de gerenciamento de dados mestres operacionais poderia ajudar a resolver essas deficiências.



**Figura 7 - Elementos principais da arquitetura Kimball DW / BI.**

### Extract, Transformation, and Load System

O sistema de extração, transformação e carga (ETL) do ambiente DW / BI consiste em uma área de trabalho, estruturas de dados instanciados e um conjunto de processos. O sistema ETL é tudo entre os sistemas de origem operacionais e a área de apresentação do DW / BI. Nós elaboramos a arquitetura de sistemas ETL e técnicas associadas no Capítulo 19: Subsistemas e Técnicas de ETL, mas queremos apresentar essa parte fundamental do quebra-cabeça geral do sistema DW / BI.

A extração é a primeira etapa no processo de obtenção de dados no ambiente de data warehouse. Extração significa ler e entender os dados de origem e copiar os dados necessários no sistema ETL para manipulação adicional. Neste ponto, os dados pertencem ao data warehouse.

Depois que os dados são extraídos para o sistema ETL, existem inúmeras transformações potenciais, como limpar os dados (corrigir erros de ortografia, resolver conflitos de domínio, lidar com elementos ausentes ou analisar em formatos padrão), combinando dados de várias fontes e -duplicando dados. O sistema ETL agrega valor aos dados com essas tarefas de limpeza e conformidade, alterando os dados e aprimorando-os. Além disso, essas atividades podem ser arquitetadas para criar metadados de diagnóstico, eventualmente levando à reengenharia de processos de negócios para melhorar a qualidade dos dados nos sistemas de origem ao longo do tempo.

A etapa final do processo de ETL é a estruturação física e o carregamento de dados nos modelos dimensionais de destino da área de apresentação. Como a principal missão do sistema ETL é entregar as tabelas de dimensões e fatos na etapa de entrega, esses subsistemas são críticos. Muitos desses subsistemas definidos concentram-se no processamento da tabela de dimensões, como atribuições de chave substituta, consultas de código para fornecer descrições apropriadas, divisão ou combinação de colunas para apresentar os valores de dados apropriados ou unir estruturas de tabela de formulário normais subjacentes em dimensões desordenadas desnormalizadas . Em contraste, as tabelas de fatos costumam ser grandes e consomem muito tempo para serem carregadas, mas prepará-las para a área de apresentação é bastante simples. Quando as tabelas de dimensão e de fato em um modelo dimensional foram atualizadas, indexadas, fornecidas com agregados apropriados e com qualidade adicional garantida, a comunidade de negócios não notou que os novos dados foram publicados.

Permanece a consternação da indústria sobre se os dados no sistema ETL devem ser reaproveitados em estruturas físicas normalizadas antes de serem carregados nas estruturas dimensionais da área de apresentação para consulta e relatório. O sistema ETL é tipicamente dominado pelas atividades simples de classificação e processamento seqüencial. Em muitos casos, o sistema ETL não é baseado em tecnologia relacional, mas pode depender de um sistema de arquivos. Após validar os dados para conformidade com as regras de negócios definidas um a um e muitos para um, pode ser inútil dar o passo final de construir um banco de dados físico 3NF, antes de transformar os dados novamente em desnormalizados. estruturas para a área de apresentação do BI.

No entanto, há casos em que os dados chegam à porta do sistema ETL em um formato relacional 3NF. Nessas situações, os desenvolvedores do sistema ETL podem se sentir mais confortáveis executando as tarefas de limpeza e transformação usando estruturas normalizadas. Embora um banco de dados normalizado para processamento de ETL seja aceitável, temos algumas reservas sobre essa abordagem. A criação de estruturas normalizadas para o ETL e estruturas dimensionais para apresentação significa que os dados são potencialmente extraídos, transformados e carregados duas vezes - uma vez no banco de dados normalizado e, novamente, quando você carrega o modelo dimensional. Obviamente, esse processo de duas etapas requer mais tempo e investimento para o desenvolvimento, mais tempo para o carregamento periódico ou atualização de dados e mais capacidade para armazenar as várias cópias dos dados. Na linha de fundo, isso normalmente se traduz na necessidade de maiores orçamentos de desenvolvimento, suporte contínuo e plataforma de hardware.

Infelizmente, algumas iniciativas de DW / BI falharam miseravelmente porque concentraram toda a sua energia e recursos na construção das estruturas normalizadas, em vez de alocar tempo para desenvolver uma área de apresentação dimensional que ofereça suporte a uma melhor tomada de decisões de negócios. Embora a consistência de dados em toda a empresa seja uma meta fundamental do ambiente DW / BI, pode haver abordagens eficazes e menos dispendiosas do que criar fisicamente tabelas normalizadas no sistema ETL, se essas estruturas ainda não existirem.

|  |
| --- |
| Nota |
| É aceitável criar um banco de dados normalizado para suportar os processos de ETL; no entanto, esse não é o objetivo final. As estruturas normalizadas devem estar fora dos limites das consultas do usuário, porque elas anulam os objetivos gêmeos de compreensibilidade e desempenho. |

### Área de apresentação para suporte de business intelligence

A área de apresentação do DW / BI é onde os dados são organizados, armazenados e disponibilizados para consultas diretas por usuários, redatores de relatórios e outros aplicativos analíticos de BI. Como o sistema ETL do back-room está fora dos limites, a área de apresentação é o ambiente DW / BI, no que diz respeito à comunidade empresarial; é tudo que a empresa vê e toca por meio de suas ferramentas de acesso e aplicativos de BI. O título original de pré-lançamento para a primeira edição do The Data Warehouse Toolkit foi Getting the Data Out. Isto é o que a área de apresentação com seus modelos dimensionais é toda sobre.

Temos várias opiniões fortes sobre a área de apresentação. Em primeiro lugar, insistimos que os dados sejam apresentados, armazenados e acessados em esquemas dimensionais, esquemas de estrela relacional ou cubos OLAP. Felizmente, o setor amadureceu até o ponto em que não estamos mais debatendo essa abordagem; concluiu que a modelagem dimensional é a técnica mais viável para entregar dados aos usuários do DW / BI.

Nossa segunda participação no terreno sobre a área de apresentação é que ela deve conter dados atômicos detalhados. Dados atômicos são necessários para resistir a ataques de consultas imprevisíveis de usuários ad hoc. Embora a área de apresentação também possa conter dados agregados que melhoram o desempenho, não é suficiente fornecer esses resumos sem os dados granulares subjacentes em uma forma dimensional. Em outras palavras, é completamente inaceitável armazenar apenas dados resumidos em modelos dimensionais enquanto os dados atômicos são bloqueados em modelos normalizados. É impraticável esperar que um usuário faça uma busca detalhada por dados dimensionais quase até o nível mais granular e, em seguida, perca os benefícios de uma apresentação dimensional na etapa final. Embora os usuários e aplicativos do DW / BI possam parecer pouco comuns em um único item de linha em um pedido, eles podem estar muito interessados ​​nos pedidos da semana anterior de produtos de um determinado tamanho (ou tipo de pacote ou fabricante) para clientes que adquiridos nos últimos 6 meses (ou residir em um determinado estado ou ter certos termos de crédito). Os dados mais detalhados devem estar disponíveis na área de apresentação para que os usuários possam fazer as perguntas mais precisas possíveis. Como os requisitos dos usuários são imprevisíveis e estão constantemente mudando, você precisa fornecer acesso aos detalhes mais sofisticados para que eles possam ser respondidos para resolver as questões do momento.

A área de dados de apresentação deve ser estruturada em torno de eventos de medição de processos de negócios. Essa abordagem se alinha naturalmente aos sistemas operacionais de captura de dados de origem. Modelos dimensionais devem corresponder a eventos de captura de dados físicos; eles não devem ser projetados para entregar o relatório do dia. Os processos de negócios de uma empresa ultrapassam os limites 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 separados semelhantes, mas ligeiramente diferentes, contendo métricas de vendas para as equipes de vendas, marketing, logística e finanças.

Todas as estruturas dimensionais devem ser construídas usando dimensões comuns e conformadas. Essa é a base da arquitetura do barramento de data warehouse corporativo descrita no Capítulo 4. A aderência à arquitetura de barramento é a participação final no terreno para a área de apresentação. Sem dimensões conformes e compartilhadas, um modelo dimensional se torna um aplicativo independente. Conjuntos de dados stovepipe isolados que não podem ser interligados são a ruína do movimento DW / BI, pois perpetuam visões incompatíveis da empresa. Se você tem alguma esperança de construir um ambiente DW / BI robusto e integrado, deve comprometer-se com a arquitetura de barramento corporativo. Quando modelos dimensionais foram projetados com dimensões conformadas, eles podem ser prontamente combinados e usados juntos. A área de apresentação em uma solução DW / BI corporativa grande consiste, em última instância, em dezenas de modelos dimensionais com muitas das tabelas de dimensões associadas compartilhadas entre as tabelas de fatos.

Usar a arquitetura de barramento é o segredo para construir sistemas DW / BI distribuídos. Quando a arquitetura de barramento é usada como uma estrutura, é possível desenvolver o data warehouse corporativo de maneira ágil, descentralizada, realisticamente delimitada e iterativa.

|  |
| --- |
| Nota |
| Os dados na área de apresentação questionável do sistema DW / BI devem ser dimensionais, atômicos (complementados por agregados que melhoram o desempenho), centrados em processos de negócios e aderir à arquitetura de barramento de data warehouse empresarial. Os dados não devem ser estruturados de acordo com a interpretação dos dados pelos departamentos individuais. |

### Aplicações de Business Intelligence

O principal componente final da arquitetura Kimball DW / BI é o aplicativo de business intelligence (BI). O termo aplicativo de BI se refere vagamente ao conjunto de recursos fornecidos aos usuários corporativos para alavancar a área de apresentação para a tomada de decisões analíticas. Por definição, todos os aplicativos de BI consultam os dados na área de apresentação do DW / BI. Consultar, obviamente, é o objetivo de usar dados para melhorar a tomada de decisões.

Um aplicativo de BI pode ser tão simples quanto uma ferramenta de consulta ad hoc ou tão complexo quanto um aplicativo sofisticado de mineração de dados ou modelagem. As ferramentas de consulta ad hoc, por mais poderosas que sejam, podem ser compreendidas e usadas efetivamente apenas por uma pequena porcentagem da população potencial de usuários de negócios DW / BI. A maioria dos usuários de negócios provavelmente acessará os dados por meio de aplicativos pré-construídos e modelos orientados por parâmetros que não exigem que os usuários criem consultas diretamente. Alguns dos aplicativos mais sofisticados, como ferramentas de modelagem ou previsão, podem carregar os resultados de volta para os sistemas de origem operacionais, sistema ETL ou área de apresentação.

### Metáfora do restaurante para a arquitetura de Kimball

Uma de nossas metáforas favoritas reforça a importância de separar o ambiente geral do DW / BI em componentes distintos. Nesse caso, consideraremos as semelhanças entre um restaurante e o ambiente DW / BI.

#### ETL na cozinha dos fundos

O sistema ETL é análogo à cozinha de um restaurante. A cozinha do restaurante é um mundo em si mesmo. Chefs talentosos pegam matérias-primas e as transformam em refeições apetitosas e deliciosas para os clientes do restaurante. Mas muito antes de uma cozinha comercial entrar em operação, uma quantidade significativa de planejamento é dedicada ao projeto do layout e dos componentes do espaço de trabalho.

A cozinha é organizada com vários objetivos de design em mente. Primeiro, o layout deve ser altamente eficiente. Os gerentes de restaurantes querem alta produtividade na cozinha. Quando o restaurante está lotado e todo mundo está com fome, não há tempo para o movimento desperdiçado. Proporcionar qualidade consistente da cozinha do restaurante é o segundo objetivo importante. O estabelecimento está condenado se as placas que saem da cozinha repetidamente falham em atender às expectativas. Para alcançar a consistência, os chefs criam seus molhos especiais uma vez na cozinha, em vez de enviar os ingredientes para a mesa, onde as variações inevitavelmente ocorrerão. Finalmente, a produção da cozinha, as refeições entregues aos clientes do restaurante, também devem ser de alta integridade. Você não gostaria que alguém enviasse intoxicação alimentar ao jantar em seu restaurante. Consequentemente, as cozinhas são projetadas com integridade em mente; A preparação da salada não acontece nas mesmas superfícies onde o frango cru é manipulado.

Assim como a qualidade, a consistência e a integridade são considerações importantes ao projetar a cozinha do restaurante, elas também são preocupações constantes para a administração cotidiana do restaurante. Chefs se esforçam para obter as melhores matérias-primas possíveis. Os produtos adquiridos devem atender aos padrões de qualidade e serão rejeitados se não atenderem aos padrões mínimos. A maioria dos restaurantes finos modifica seus cardápios com base na disponibilidade de ingredientes de qualidade.

A equipe do restaurante tem sua cozinha com profissionais qualificados empunhando as ferramentas de seu ofício. Cozinheiros manipulam facas afiadas com incrível confiança e facilidade. Eles operam equipamentos poderosos e trabalham em superfícies extremamente quentes sem incidentes.

Dado o ambiente perigoso, a cozinha do quarto dos fundos está fora dos limites para os clientes do restaurante. Coisas acontecem na cozinha que os clientes simplesmente não deveriam ver. Simplesmente não é seguro. Cozinheiros profissionais que manipulam facas afiadas não devem se distrair com as perguntas dos clientes. Você também não gostaria que os clientes entrassem na cozinha para mergulhar os dedos em um molho para ver se eles queriam pedir uma entrada. Para evitar essas intrusões, a maioria dos restaurantes tem uma porta fechada que separa a cozinha da área onde os comensais são servidos. Mesmo os restaurantes que possuem um formato de cozinha aberta normalmente têm uma barreira, como uma parede parcial de vidro, separando os dois ambientes. Os clientes são convidados a assistir, mas não podem entrar na cozinha. Embora parte da cozinha possa ser visível, há sempre salas de apoio fora de vista onde a preparação menos visualmente desejável ocorre.

O sistema de ETL do data warehouse se assemelha à cozinha do restaurante. Os dados da fonte são magicamente transformados em informações significativas e apresentáveis. O sistema ETL da sala dos fundos deve ser organizado e arquitetado muito antes de qualquer dado ser extraído da fonte. Como a cozinha, o sistema ETL é projetado para garantir rendimento. Ele deve transformar os dados da fonte bruta no modelo de destino com eficiência, minimizando o movimento desnecessário.

Obviamente, o sistema ETL também está altamente preocupado com a qualidade, integridade e consistência dos dados. Os dados recebidos são verificados quanto à qualidade razoável conforme eles entram. As condições são monitoradas continuamente para garantir que as saídas de ETL sejam de alta integridade. As regras de negócios para derivar consistentemente métricas e atributos de valor agregado são aplicadas uma vez por profissionais qualificados no sistema ETL, em vez de depender de cada patrono para desenvolvê-los independentemente. Sim, isso sobrecarrega a equipe de ETL, mas é feito para oferecer um produto melhor e mais consistente aos usuários do DW / BI.

|  |
| --- |
| Nota |
| Um ambiente DW / BI adequadamente projetada troca o trabalho nas aplicações de BI da sala frontal em favor do trabalho no sistema ETL da sala dos fundos. O trabalho na sala da frente deve ser feito repetidamente pelos usuários corporativos, enquanto o trabalho na sala de apoio é feito uma vez pela equipe da ETL. |

Finalmente, o sistema ETL deve estar fora dos limites para os usuários corporativos e desenvolvedores de aplicativos de BI. Assim como você não quer que os clientes do restaurante andem pela cozinha e consumam alimentos semi-cozidos, você não quer profissionais ocupados em ETL distraídos por perguntas imprevisíveis de usuários de BI. As conseqüências podem ser altamente desagradáveis se os usuários mergulharem seus dedos em potes intermediários enquanto a preparação de dados ainda estiver em andamento. Assim como na cozinha do restaurante, ocorrem atividades no sistema ETL que os usuários do DW / BI não devem ver. Quando os dados estão prontos e a qualidade é verificada para o consumo do usuário, eles são levados até a área de apresentação do DW / BI.

#### Apresentação de dados e BI na sala de jantar da frente

Agora, volte sua atenção para a sala de jantar do restaurante. Quais são os principais fatores que diferenciam os restaurantes? De acordo com as classificações e resenhas populares de restaurantes, os restaurantes são tipicamente pontuados em quatro qualidades distintas:

* Comida (qualidade, sabor e apresentação)
* Decoração (ambiente atraente e confortável para os clientes)
* Serviço (pronta entrega de comida, pessoal de apoio atento e comida recebida como ordenado)
* Custo

A maioria dos clientes se concentra inicialmente na pontuação dos alimentos quando avaliam as opções de refeições. Em primeiro lugar, o restaurante serve boa comida? Esse é o produto principal do restaurante. No entanto, os fatores de decoração, serviço e custo também afetam a experiência gastronômica geral dos clientes e são considerações ao avaliar se comer em um restaurante.

Naturalmente, a entrega principal da cozinha DW / BI são os dados na área de apresentação. Quais dados estão disponíveis? Como o restaurante, o sistema DW / BI fornece "menus" para descrever o que está disponível por meio de metadados, relatórios publicados e aplicativos analíticos com parâmetros. Os clientes do DW / BI esperam consistência e alta qualidade. Os dados da área de apresentação devem estar devidamente preparados e seguros para consumo.

A decoração da área de apresentação deve ser organizada para o conforto dos clientes. Deve ser projetado com base nas preferências dos clientes do BI e não na equipe de desenvolvimento. O serviço também é crítico no sistema DW / BI. Os dados devem ser entregues, conforme solicitado, prontamente em um formato atraente para o usuário corporativo ou desenvolvedor de aplicativos de BI.

Finalmente, o custo é um fator para o sistema DW / BI. O pessoal da cozinha pode estar a sonhar com refeições caras e elaboradas, mas se não houver mercado a esse preço, o restaurante não sobreviverá.

Se os clientes do restaurante gostam da sua experiência de jantar, então tudo é róseo

o gerente do restaurante. A sala de jantar está sempre ocupada; às vezes há até uma lista de espera. As métricas de desempenho do gerente do restaurante são todas promissoras: números elevados de comensais, turnovers de mesa e receita e lucro noturnos, enquanto a rotatividade de pessoal é baixa. As coisas parecem tão boas que o dono do restaurante está considerando um site de expansão para lidar com o tráfego. Por outro lado, se os clientes do restaurante não estiverem felizes, as coisas desmoronam rapidamente. Com um número limitado de clientes, o restaurante não está ganhando dinheiro suficiente para cobrir suas despesas, e a equipe não está dando gorjetas. Em um tempo relativamente curto, o restaurante fecha.

Os gerentes de restaurantes costumam verificar proativamente a satisfação de seus clientes com a comida e a experiência gastronômica. Se um patrono está insatisfeito, eles tomam medidas imediatas para corrigir a situação. Da mesma forma, os gerentes de DW / BI devem monitorar proativamente a satisfação. Você não pode pedir a espera para ouvir reclamações. Muitas vezes, as pessoas abandonam um restaurante sem expressar suas preocupações. Com o tempo, os gerentes percebem que as contas de restaurantes caíram, mas talvez nem saibam por quê.

Inevitavelmente, os patronos anteriores do DW / BI localizarão outro “restaurante” que melhor se adapte às suas necessidades e preferências, desperdiçando os milhões de dólares investidos para projetar, construir e empregar o sistema DW / BI. Claro, você pode evitar esse final infeliz gerenciando o restaurante de forma proativa; certifique-se de que a cozinha esteja organizada e utilizada adequadamente para entregar conforme necessário à comida, decoração, serviço e custo da área de apresentação.

## Arquiteturas alternativas DW / BI

Tendo acabado de descrever a arquitetura Kimball, vamos discutir várias outras abordagens arquiteturais do DW / BI. Vamos rever rapidamente as duas alternativas dominantes para a arquitetura Kimball, destacando as semelhanças e diferenças. Em seguida, fecharemos esta seção concentrando-nos em uma abordagem híbrida que combine alternativas.

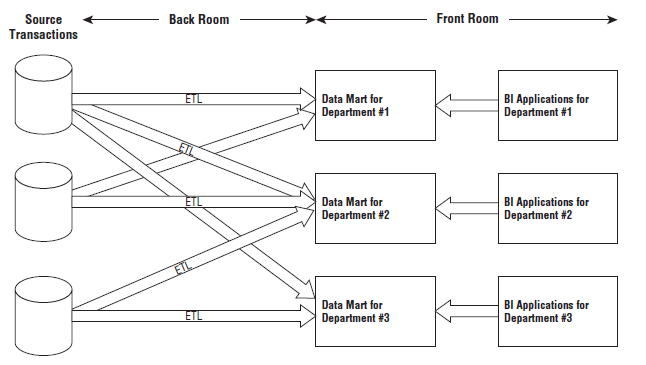
Felizmente, nas últimas décadas, as diferenças entre a arquitetura de Kimball e as alternativas foram atenuadas. Ainda mais feliz, existe um papel para a modelagem dimensional, independentemente da sua predisposição arquitetônica.

Reconhecemos que as organizações construíram com sucesso sistemas DW / BI com base nas abordagens defendidas por outros. Acreditamos firmemente que, em vez de encorajar mais consternação sobre nossas diferenças filosóficas, seria muito melhor para a indústria dedicar energia para garantir que nossas entregas de DW / BI sejam amplamente aceitas pela empresa para tomar decisões melhores e mais informadas. A arquitetura deve ser apenas um meio para esse objetivo.

### Arquitetura de Data Mart Independente

Com essa abordagem, os dados analíticos são implantados em uma base departamental sem a preocupação de compartilhar e integrar informações em toda a empresa, conforme ilustrado na Figura 1-8. Normalmente, um único departamento identifica requisitos para dados de um sistema de origem operacional. O departamento trabalha com a equipe de TI ou consultores externos para construir um banco de dados que satisfaça suas necessidades departamentais, refletindo suas regras de negócios e rotulagem preferida. Trabalhando isoladamente, esse datamart departamental atende aos requisitos analíticos do departamento.

Enquanto isso, outro departamento está interessado nos mesmos dados de origem. É extremamente comum que vários departamentos se interessem pelas mesmas métricas de desempenho resultantes dos principais eventos do processo de negócios de uma organização. Mas como esse departamento não tem acesso ao data mart inicialmente construído pelo outro departamento, ele segue um caminho semelhante por si próprio, obtendo recursos e criando uma solução departamental que contém dados semelhantes, mas ligeiramente diferentes. Quando os usuários de negócios desses dois departamentos discutem o desempenho organizacional com base nos relatórios de seus respectivos repositórios, não surpreende que nenhum dos números corresponda devido às diferenças nas regras de negócios e na rotulagem.



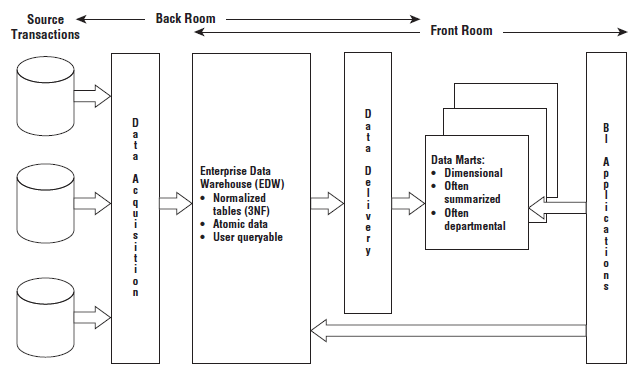
**Figura 8 - Ilustração simplificada da “arquitetura” do data mart independente.**

Esses silos analíticos autônomos representam uma "arquitetura" de DW / BI que é essencialmente não arquitetada. Embora nenhum líder do setor defenda esses data marts independentes, essa abordagem é predominante, especialmente em grandes organizações. Ele reflete o modo como muitas organizações financiam projetos de TI, além de não exigir governança e coordenação de dados entre organizações. É o caminho de menor resistência para desenvolvimento rápido a um custo relativamente baixo, pelo menos a curto prazo. Naturalmente, múltiplos extratos descoordenados das mesmas fontes operacionais e armazenamento redundante de dados analíticos são ineficientes e inúteis a longo prazo. Sem qualquer perspectiva corporativa, essa abordagem independente resulta em uma miríade de soluções autônomas que perpetuam visões incompatíveis do desempenho da organização, resultando em debates e reconciliações organizacionais desnecessários.

Nós desencorajamos fortemente a abordagem de data mart independente. No entanto, muitas vezes esses data marts independentes adotaram a modelagem dimensional porque estão interessados em fornecer dados que sejam fáceis para a empresa entender e responder às consultas. Portanto, nossos conceitos de modelagem dimensional são frequentemente aplicados a essa arquitetura, apesar da total desconsideração de alguns de nossos princípios fundamentais, como foco em detalhes atômicos, construção por processo de negócios em vez de departamento e alavancagem de dimensões conformadas para consistência e integração da empresa.

### Fábrica de informações corporativas de hub e spor Inmon Architecture

A abordagem CIF (Corporate Information Factory - Fábrica de Informações Corporativas) centrada e defendida é defendida por Bill Inmon e outros profissionais do setor. A Figura 1-9 ilustra uma versão simplificada do CIF, enfocando os principais elementos e conceitos que justificam a discussão.



**Figura 9 - Ilustração simplificada da fábrica de informações corporativas hub-and-spoke**

Com o CIF, os dados são extraídos dos sistemas de origem operacionais e processados por meio de um sistema ETL, às vezes chamado de aquisição de dados. Os dados atômicos resultantes desse processamento são armazenados em um banco de dados 3NF; esse repositório atômico normalizado é chamado de Enterprise Data Warehouse (EDW) dentro da arquitetura CIF. Embora a arquitetura Kimball permita a normalização opcional para suportar o processamento de ETL, o EDW normalizado é uma construção obrigatória no CIF. Como a abordagem Kimball, o CIF defende a coordenação e integração de dados corporativos. O CIF diz que o EDW normalizado preenche essa função, enquanto a arquitetura Kimball enfatiza a importância de um barramento corporativo com dimensões conformadas.

|  |
| --- |
| Nota |
| O processo de normalização não fala tecnicamente com integração. A normalização simplesmente cria tabelas físicas que implementam relacionamentos de muitos para um. A integração, por outro lado, exige que as inconsistências decorrentes de fontes separadas sejam resolvidas. Fontes de bancos de dados incompatíveis separadas podem ser normalizadas ao máximo sem abordar a integração. A arquitetura Kimball baseada em dimensões conformes inverte essa lógica e se concentra na resolução de inconsistências de dados sem exigir explicitamente a normalização. |

As organizações que adotaram a abordagem CIF geralmente têm usuários de negócios acessando o repositório EDW devido a seu nível de detalhes ou disponibilidade de dados. No entanto, os processos subseqüentes de entrega de dados ETL também preenchem os ambientes analíticos e de relatórios downstream para dar suporte aos usuários corporativos. Embora muitas vezes estruturadas dimensionalmente, os bancos de dados analíticos resultantes normalmente diferem das estruturas na área de apresentação da arquitetura Kimball, pois são frequentemente centrados no departamento (em vez de organizados em torno de processos de negócios) e preenchidos com dados agregados (em vez de detalhes atômicos). Se os processos de ETL de entrega de dados aplicam regras de negócios além da sumarização básica, como renomeação departamental de colunas ou cálculos alternativos, pode ser difícil vincular esses bancos de dados analíticos ao repositório atômico do EDW.

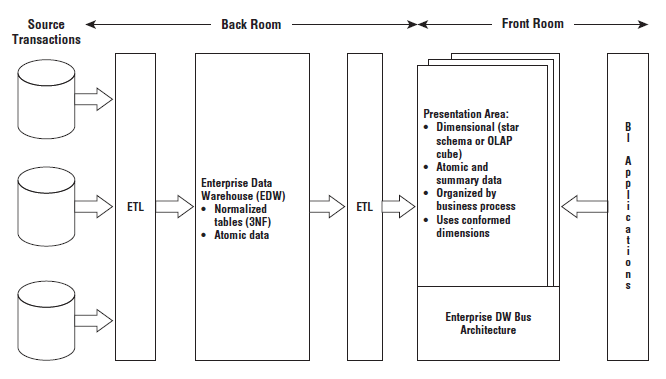
|  |
| --- |
| Nota |
| A forma mais extrema de uma arquitetura CIF pura é impraticável como um data warehouse, em nossa opinião. Essa arquitetura bloqueia os dados atômicos em estruturas normalizadas de difícil consulta, ao mesmo tempo em que fornece data marts incompatíveis com departamentos para diferentes grupos de usuários corporativos. Mas antes de ficar deprimido demais com essa visão, fique atento para a próxima seção. |

### Arquitetura Híbrida Hub-and-Spoke e Kimball

A arquitetura final que justifica a discussão é o casamento das arquiteturas Kimball e Inmon CIF. Conforme ilustrado na Figura 1-10, essa arquitetura preenche um EDW centrado em CIF que está completamente fora dos limites para usuários de negócios para análise e geração de relatórios. É apenas a fonte para preencher uma área de apresentação no estilo Kimball na qual os dados são dimensionais, atômicos (complementados por agregados), centrados em processos e em conformidade com a arquitetura de barramento de data warehouse corporativo.

Alguns proponentes dessa abordagem combinada afirmam que é o melhor dos dois mundos. Sim, combina as duas abordagens orientadas para a empresa. Ele pode alavancar um investimento preexistente em um repositório integrado, ao mesmo tempo em que aborda os problemas de desempenho e usabilidade associados ao 3NF EDW, enviando consultas para a área de apresentação dimensional. E como a entrega final para os usuários de negócios e aplicativos de BI é construída com base nos princípios de Kimball, quem pode argumentar com a abordagem?

Se você já investiu na criação de um 3NF EDW, mas não está atendendo às expectativas dos usuários de relatórios e análises rápidos e flexíveis, essa abordagem híbrida pode ser apropriada para sua organização. Se você estiver começando com uma folha de papel em branco, a abordagem híbrida provavelmente custará mais tempo e dinheiro, tanto durante o desenvolvimento quanto na operação contínua, dados os múltiplos movimentos de dados e o armazenamento redundante de detalhes atômicos. Se você tiver o apetite, a necessidade percebida e talvez o mais importante, o orçamento e a paciência organizacional para normalizar completamente e instanciar seus dados antes de carregá-los em estruturas dimensionais bem projetadas de acordo com os métodos de Kimball, vá em frente.



**Figura 10 - Arquitetura híbrida com estruturas 3NF e área de apresentação dimensional de Kimball.**

## Mitos de Modelagem Dimensional

Apesar da ampla aceitação da modelagem dimensional, algumas percepções errôneas persistem na indústria. Essas falsas afirmações são uma distração, especialmente quando você deseja alinhar sua equipe com práticas recomendadas comuns. Se as pessoas na sua organização continuamente criticam a modelagem dimensional, esta seção deve estar em sua lista de leitura recomendada; suas percepções podem ser obscurecidas por esses mal-entendidos comuns.

### Mito 1: Modelos dimensionais são apenas para dados de resumo

Este primeiro mito é freqüentemente a causa básica de modelos dimensionais mal projetados. Como você não pode prever todas as perguntas feitas pelos usuários corporativos, é necessário fornecer acesso questionável aos dados mais detalhados para que eles possam ser atualizados com base na pergunta da empresa. Dados no mais baixo nível de detalhe são praticamente impermeáveis a surpresas ou mudanças. Os dados de resumo devem complementar os detalhes granulares apenas para fornecer desempenho aprimorado para consultas comuns, mas não para substituir os detalhes.

Um corolário relacionado a este primeiro mito é que apenas uma quantidade limitada de dados históricos deve ser armazenada em estruturas dimensionais. Nada sobre um modelo dimensional proíbe armazenar histórico substancial. A quantidade de histórico disponível em modelos dimensionais só deve ser orientada pelos requisitos da empresa.

### Mito 2: Modelos dimensionais são departamentais, não corporativos

Em vez de traçar limites com base nos departamentos organizacionais, os modelos dimensionais devem ser organizados em torno dos processos de negócios, como pedidos, faturas e chamadas de serviço. Várias funções de negócios geralmente desejam analisar as mesmas métricas resultantes de um único processo de negócios. Vários extratos dos mesmos dados de origem que criam bancos de dados analíticos múltiplos e inconsistentes devem ser evitados.

### Mito 3: Modelos dimensionais não são escalonáveis

Modelos dimensionais são extremamente escalonáveis. Tabelas de fatos freqüentemente têm bilhões de linhas; Tabelas de fatos contendo 2 trilhões de linhas foram relatadas. Os fornecedores de banco de dados abraçaram totalmente o DW / BI e continuam a incorporar recursos em seus produtos para otimizar a escalabilidade e o desempenho dos modelos dimensionais. Os modelos normalizados e dimensionais contêm os mesmos relacionamentos de informação e dados; o conteúdo lógico é idêntico. Cada relação de dados expressa em um modelo pode ser expressa com precisão no outro. Modelos normalizados e dimensionais podem responder exatamente as mesmas questões, embora com dificuldade variável.

### Mito 4: Modelos dimensionais são apenas para uso previsível

Os modelos dimensionais não devem ser projetados com foco em relatórios ou análises pré-definidos; o design deve centrar-se nos processos de medição. Obviamente, é importante considerar os requisitos de filtragem e filtragem da aplicação de BI. Mas você não deve criar uma lista dos dez principais relatórios no vácuo porque essa lista está fadada a mudar, tornando o modelo dimensional um alvo em movimento. A chave é se concentrar nos eventos de medição da organização que são normalmente estáveis, ao contrário das análises que estão em constante evolução. tabelas de fatos no nível mais granular. Modelos dimensionais que fornecem apenas dados de resumo estão fadados a ser problemáticos; os usuários correm em paredes analíticas de tijolos quando tentam detalhar detalhes que não estão disponíveis nas tabelas de resumo. Os desenvolvedores também se deparam com paredes de tijolos porque não podem acomodar facilmente novas dimensões, atributos ou fatos com essas tabelas resumidas prematuramente. O ponto de partida correto para seus modelos dimensionais é expressar os dados com o menor detalhe possível para o máximo

flexibilidade e extensibilidade. Lembre-se de que, ao pré-supor a questão comercial, você provavelmente pré-resumirá os dados, o que pode ser fatal no longo prazo. Como credita o arquiteto Mies van der Rohe, “Deus está nos detalhes”. Fornecer modelos dimensionais preenchidos com os dados mais detalhados possíveis garante máxima flexibilidade e extensibilidade. Entregar menos em seus modelos dimensionais prejudica a base necessária para uma inteligência de negócios robusta

### Mito 5: Modelos dimensionais não podem ser integrados

Modelos dimensionais certamente podem ser integrados se estiverem em conformidade com a arquitetura de barramento de data warehouse empresarial. As dimensões conformadas são criadas e mantidas como dados mestre centralizados e persistentes no sistema ETL e, em seguida, reutilizadas em modelos dimensionais para permitir a integração de dados e garantir consistência semântica. A integração de dados depende de rótulos, valores e definições padronizados. É difícil chegar a um consenso organizacional e implementar as regras de ETL correspondentes, mas você não pode evitar o efeito, independentemente de estar preenchendo modelos normalizados ou dimensionais. Os bancos de dados da área de apresentação que não aderem à arquitetura de barramento com dimensões conformadas compartilhadas levam a soluções autônomas. Você não pode responsabilizar a modelagem dimensional pelo fracasso das organizações em adotar um de seus princípios fundamentais.

## Mais razões para pensar dimensionalmente

A maior parte deste livro se concentra na modelagem dimensional para projetar bancos de dados na área de apresentação do DW / BI. Mas os conceitos de modelagem dimensional vão além do design de estruturas de dados simples e rápidas. Você deve pensar dimensionalmente em outros momentos críticos de um projeto DW / BI.

Ao reunir requisitos para uma iniciativa de DW / BI, você precisa ouvir e sintetizar os resultados dos processos de negócios. Às vezes, as equipes são levadas a se concentrar em um conjunto de relatórios necessários ou indicadores de painel. Em vez disso, você deve se perguntar constantemente sobre os eventos de medição do processo de negócios que produzem o relatório ou as métricas do painel. Ao especificar o escopo do projeto, você deve se manter firme para se concentrar em um único processo de negócios por projeto e não se inscrever para implantar um painel que cubra um punhado deles em uma única iteração.

Embora seja essencial que a equipe de DW / BI se concentre nos processos de negócios, é igualmente importante obter o gerenciamento de TI e de negócios no mesmo comprimento de onda. Devido às políticas históricas de financiamento de TI, o negócio pode estar mais familiarizado com implantações de dados departamentais. Você precisa mudar sua mentalidade sobre o lançamento do DW / BI para uma perspectiva de processo. Ao priorizar oportunidades e desenvolver o roteiro do DW / BI, os processos de negócios são a unidade de trabalho. Felizmente, o gerenciamento de negócios normalmente adota essa abordagem porque reflete seu pensamento sobre os principais indicadores de desempenho. Além disso, eles viveram com as inconsistências, debates incessantes e intermináveis ​​reconciliações causadas pela abordagem departamental, então eles estão prontos para uma nova tática. Trabalhando com parceiros de liderança de negócios, classifique cada processo de negócios em valor de negócio e viabilidade e, em seguida, resolva os processos com os maiores escores de impacto e viabilidade primeiro. Embora a priorização seja uma atividade conjunta com a empresa, sua compreensão subjacente dos processos de negócios da organização é essencial para sua eficácia e subsequente capacidade de execução.

Se você tiver a tarefa de projetar a arquitetura de dados do sistema DW / BI, será necessário envolver os processos da organização, juntamente com os dados da dimensão descritiva principal associados. A entrega principal para esta atividade, a matriz de barramento de data warehouse empresarial, será totalmente verificada no Capítulo 4. A matriz também serve como uma ferramenta útil para divulgar os benefícios potenciais de uma plataforma de gerenciamento de dados mestres mais rigorosa.

Os programas de administração ou governança de dados devem se concentrar primeiro nas principais dimensões. Dependendo do setor, a lista pode incluir data, cliente, produto, funcionário, instalação, provedor, aluno, corpo docente, conta 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 da comunidade empresarial. Estabelecer as responsabilidades de governança de dados para esses nomes é a chave para implantar, eventualmente, dimensões que forneçam consistência e atendam às necessidades da empresa de filtragem analítica, agrupamento e rotulagem. Dimensões robustas se traduzem em sistemas robustos DW / BI.

Como você pode ver, a motivação fundamental para a modelagem dimensional é frontal e central muito antes de projetar esquemas em estrela ou cubos OLAP. Da mesma forma, o modelo dimensional permanecerá na vanguarda durante o sistema ETL subsequente e os designs de aplicativos de BI. Os conceitos de modelagem dimensional vinculam as comunidades comercial e técnica enquanto projetam em conjunto as entregas do DW / BI. Vamos elaborar essas ideias no Capítulo 17: Visão geral do ciclo de vida de Kimball DW / BI e no Capítulo 18: Processo de modelagem dimensional e tarefas, mas queríamos plantar as sementes cedo para que elas tivessem tempo de germinar.

## Considerações ágeis

Atualmente, há interesse significativo na indústria de DW / BI em práticas de desenvolvimento ágil. Com o risco de simplificação excessiva, as metodologias ágeis concentram-se em incrementos dimensionáveis de trabalho que podem ser concluídos dentro de prazos razoáveis medidos em semanas, em vez de abordar um projeto de escopo muito maior (e, portanto, mais arriscado) com resultados prometidos em meses ou anos. Soa bem, não é?

Muitos dos princípios básicos das metodologias ágeis estão alinhados com as melhores práticas da Kimball, incluindo

* Concentre-se no fornecimento de valor comercial. Este tem sido o mantra Kimball por décadas.
* Valorize a colaboração entre a equipe de desenvolvimento e as partes interessadas do negócio. Como o campo ágil, incentivamos fortemente uma parceria próxima com o negócio.
* Estresse a comunicação, feedback e priorização presenciais em andamento com as partes interessadas da empresa.
* Adaptar-se rapidamente a requisitos inevitavelmente em evolução.
* Enfrentar o desenvolvimento de maneira iterativa e incremental.

Embora essa lista seja convincente, uma crítica comum às abordagens ágeis é a falta de planejamento e arquitetura, aliada a desafios contínuos de governança. A matriz de barramento de data warehouse empresarial é uma ferramenta poderosa para solucionar essas deficiências. A matriz de barramento fornece uma estrutura e um plano mestre para o desenvolvimento ágil, além de identificar as dimensões descritivas comuns reutilizáveis que fornecem consistência de dados e redução da entrega de tempo para colocação no mercado. Com o mix colaborativo certo de partes interessadas de negócios e de TI em uma sala, a matriz de barramento de data warehouse empresarial pode ser produzida em um período relativamente curto. O trabalho de desenvolvimento incremental pode produzir componentes da estrutura até que uma funcionalidade suficiente esteja disponível e depois liberada para a comunidade de negócios.

Alguns clientes e estudantes lamentam que, embora eles queiram entregar dimensões conformes consistentemente definidas em seus ambientes DW / BI, isso é “simplesmente inviável”. Eles explicam que sim, se pudessem, mas com o foco em técnicas de desenvolvimento ágil, é “impossível ”Ter tempo para obter um acordo organizacional em dimensões conformes. Argumentamos que as dimensões conformadas permitem o desenvolvimento ágil de DW / BI, juntamente com a tomada de decisões ágil. À medida que você expila o portfólio de dimensões mestras, a manivela de desenvolvimento começa a girar cada vez mais rápido. O tempo de colocação no mercado para uma nova fonte de dados de processos de negócios encolhe à medida que os desenvolvedores reutilizam as dimensões conformes existentes. Por fim, o novo desenvolvimento de ETL se concentra quase exclusivamente na entrega de mais tabelas de fatos, porque as tabelas de dimensões associadas já estão prontas para serem colocadas na prateleira.

Sem uma estrutura como a matriz de barramento de data warehouse corporativo, algumas equipes de DW / BI caíram na armadilha de usar técnicas ágeis para criar soluções analíticas ou de relatório no vácuo. Na maioria das situações, a equipe trabalhou 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 uma fonte de dados independente que outras pessoas não podem aproveitar ou, ainda pior, fornece dados que não se vinculam a outras informações analíticas da organização. Incentivamos a agilidade, quando apropriado, no entanto, a construção de conjuntos de dados isolados deve ser evitada. Tal como acontece com a maioria das coisas na vida, moderação e equilíbrio entre os extremos é quase sempre prudente.

## Resumo

Neste capítulo, discutimos as metas primordiais para os sistemas DW / BI e os conceitos fundamentais de modelagem dimensional. A arquitetura Kimball DW / BI e várias alternativas foram comparadas. Fechamos o capítulo identificando mal-entendidos comuns que alguns ainda mantêm sobre modelagem dimensional, apesar de sua ampla aceitação em todo o setor, e desafiamos você a pensar dimensionalmente além da modelagem de dados. No próximo capítulo, você fará um tour de padrões e técnicas de modelagem dimensional e, em seguida, começará a colocar esses conceitos em prática no seu primeiro estudo de caso no Capítulo 3.