**Capítulo 1**

**Fundamentos de Data Warehousing, Business Intelligence e Modelagem Dimensional**

Este primeiro capítulo estabelece a base para os capítulos seguintes. Começamos considerando os sistemas de Data Warehousing e Business Intelligence (DW/BI) de uma perspectiva de alto nível. Você pode se decepcionar ao saber que não começamos com tecnologia e ferramentas — em primeiro lugar, o sistema DW/BI deve considerar as necessidades do negócio. Com as necessidades comerciais bem definidas, trabalhamos de forma regressiva através dos designs lógicos e depois físicos, juntamente com decisões sobre tecnologia e ferramentas.

Neste capítulo, estabelecemos metas claras para o Data Warehousing e Business Intelligence, enquanto observamos as incríveis semelhanças entre as responsabilidades de um gerente de DW/BI e as de um editor.

Com essa visão ampla, exploramos os conceitos fundamentais da modelagem dimensional e estabelecemos um vocabulário básico. A partir daí, discutimos os principais componentes da arquitetura DW/BI de Kimball, junto com uma comparação de abordagens arquitetônicas alternativas; felizmente, há um papel para a modelagem dimensional, independentemente de sua inclinação arquitetônica. Finalmente, revisamos os mitos comuns sobre modelagem dimensional.

Ao final deste capítulo, você terá uma apreciação pela necessidade de ser meio administrador de banco de dados (DBA) e meio analista de negócios (MBA) ao enfrentar seu projeto DW/BI.  
O Capítulo 1 discute os seguintes conceitos:

■ Metas orientadas ao negócio de Data Warehousing e Business Intelligence  
■ Metáfora editorial para sistemas DW/BI  
■ Conceitos e vocabulário centrais da modelagem dimensional, incluindo tabelas de fatos e dimensões  
■ Componentes e princípios da arquitetura DW/BI de Kimball  
■ Comparação de arquiteturas DW/BI alternativas e o papel da modelagem dimensional em cada uma  
■ 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: registro operacional e tomada de decisão analítica. Falando simplesmente, os sistemas operacionais são onde você insere os dados, e o sistema DW/BI é onde você extrai os dados.

Os usuários de um sistema operacional movimentam as engrenagens da organização. Eles recebem pedidos, inscrevem 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 de cada vez. Eles realizam as mesmas tarefas operacionais repetidamente, executando os processos de negócios da organização. Dado esse foco na execução, os sistemas operacionais tipicamente não mantêm histórico, mas atualizam os dados para refletir o estado mais atual.

Por outro lado, os usuários de um sistema DW/BI observam as engrenagens da organização para avaliar o desempenho. Eles contam os novos pedidos e os comparam com os pedidos da semana passada, e perguntam por que os novos clientes se inscreveram e sobre o que os clientes reclamaram. Eles se preocupam em saber se os processos operacionais estão funcionando corretamente. Embora precisem de dados detalhados para apoiar suas perguntas constantemente mutáveis, os usuários do DW/BI quase nunca lidam com uma transação por vez. Esses sistemas são otimizados para consultas de alto desempenho, pois as perguntas dos usuários frequentemente requerem que centenas ou centenas de milhares de transações sejam pesquisadas e comprimidas em um conjunto de respostas. Para complicar ainda mais, os usuários de um sistema DW/BI geralmente exigem que o contexto histórico seja preservado para avaliar com precisão o desempenho da organização ao longo do tempo.

Na primeira edição de The Data Warehouse Toolkit (Wiley, 1996), Ralph Kimball dedicou um capítulo inteiro para descrever a dicotomia entre os mundos do processamento operacional e do armazenamento de dados. Atualmente, é amplamente reconhecido que o sistema DW/BI tem necessidades, clientes, estruturas e ritmos profundamente diferentes dos 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 atender à necessidade de isolar os ambientes operacionais e analíticos por razões de desempenho, eles não fazem nada para abordar as outras diferenças inerentes entre os dois tipos de sistemas. Os usuários de negócios ficam desapontados com a usabilidade e o desempenho fornecidos por esses pseudoarmazéns de dados; esses impostores prestam um desserviço ao DW/BI porque não reconhecem que seus usuários têm necessidades drasticamente diferentes dos usuários dos sistemas operacionais.

**Objetivos do Data Warehousing e Business Intelligence**

Antes de nos aprofundarmos nos detalhes da modelagem dimensional, é útil focar nos objetivos fundamentais do data warehousing e business intelligence. Os objetivos podem ser facilmente desenvolvidos ao caminhar pelos corredores de qualquer organização e ouvir a gestão empresarial.

Estes temas recorrentes existem há mais de três décadas:

■ “Coletamos toneladas de dados, mas não conseguimos acessá-los.”  
■ “Precisamos fatiar e analisar os dados de todas as formas possíveis.”  
■ “Os profissionais de negócios precisam acessar os dados facilmente.”  
■ “Só me mostre o que é importante.”  
■ “Passamos reuniões inteiras discutindo quem tem os números corretos em vez de tomar decisões.”  
■ “Queremos que as pessoas usem informações para apoiar a tomada de decisões baseada em fatos.”  
Baseando-nos em nossa experiência, essas preocupações ainda são tão universais que fundamentam os requisitos essenciais para o sistema DW/BI. Agora, transforme essas citações da gestão empresarial em requisitos.  
■ O sistema DW/BI deve tornar as informações facilmente acessíveis. Os conteúdos do sistema DW/BI devem ser compreensíveis. Os dados devem ser intuitivos e óbvios para o usuário de negócios, e não apenas para o desenvolvedor. As estruturas e rótulos dos dados devem imitar os processos de pensamento e o vocabulário dos usuários de negócios. Os usuários de negócios querem separar e combinar dados analíticos em combinações infinitas. As ferramentas de business intelligence e aplicações que acessam os dados devem ser simples e fáceis de usar. Elas também devem retornar os resultados das consultas para o usuário com tempos de espera mínimos. Podemos resumir este requisito simplesmente como simples e rápido.  
■ O sistema DW/BI deve apresentar informações de forma consistente. Os dados no sistema DW/BI devem ser credíveis. Os dados devem ser cuidadosamente montados a partir de uma variedade de fontes, limpos, com qualidade assegurada e liberados apenas quando estiverem prontos para consumo do usuário. Consistência também implica que rótulos e definições comuns para o conteúdo do sistema DW/BI sejam usados em todas as fontes de dados. Se duas medidas de desempenho têm o mesmo nome, elas devem significar a mesma coisa. Inversamente, se duas medidas não significam a mesma coisa, elas devem ser rotuladas de maneira diferente.  
■ O sistema DW/BI deve se adaptar às mudanças. As necessidades dos usuários, condições de negócios, dados e tecnologia estão todos sujeitos a mudanças. O sistema DW/BI deve ser projetado para lidar com essa mudança inevitável de maneira elegante, de modo que não invalide dados ou aplicações existentes. Dados e aplicações existentes não devem ser alterados ou interrompidos quando a comunidade de negócios fizer novas perguntas ou novos dados forem adicionados ao armazém. Finalmente, se dados descritivos no sistema DW/BI precisarem ser modificados, você deve contabilizar adequadamente as mudanças e tornar essas mudanças transparentes para os usuários.  
■ O sistema DW/BI deve apresentar informações de maneira oportuna. À medida que o sistema DW/BI é usado mais intensamente para decisões operacionais, os dados brutos podem precisar ser convertidos em informações acionáveis dentro de horas, minutos ou até segundos. A equipe DW/BI e os usuários de negócios precisam ter expectativas realistas sobre o que significa entregar dados quando há pouco tempo para limpar ou validar.  
■ O sistema DW/BI deve ser uma fortaleza segura que protege os ativos de informação. As joias da coroa informativa de uma organização estão armazenadas no data warehouse. No mínimo, o armazém provavelmente contém informações sobre o que você está vendendo, para quem 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 a base autoritária e confiável para a melhoria da tomada de decisões. O data warehouse deve ter os dados certos para apoiar a tomada de decisões. Os resultados mais importantes de um sistema DW/BI são as decisões que são tomadas com base nas evidências analíticas apresentadas; essas decisões fornecem o impacto e o valor comercial 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 de negócios deve aceitar o sistema DW/BI para considerá-lo bem-sucedido. Não importa que você tenha construído uma solução elegante usando produtos e plataformas de primeira linha. 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 onde os usuários de negócios não têm escolha a não ser usar o novo sistema, o uso do DW/BI às vezes é opcional. Os usuários de negócios abraçarão o sistema DW/BI se ele for a fonte "simples e rápida" para informações acionáveis.

Embora cada requisito nesta lista seja importante, os dois últimos são os mais críticos e, infelizmente, muitas vezes os mais negligenciados. O sucesso do data warehousing e business intelligence exige mais do que ser um arquiteto estelar, técnico, modelador ou administrador de banco de dados. Com uma iniciativa DW/BI, você tem um pé na sua zona de conforto de tecnologia da informação (TI), enquanto o outro pé está no terreno desconhecido dos usuários de negócios.

Você deve transitar entre os dois, modificando algumas habilidades comprovadas para se adaptar às demandas únicas do DW/BI. Claramente, você precisa trazer um espectro de habilidades para a festa para se comportar como se fosse um híbrido de DBA/MBA.

**Metáfora Editorial para Gerentes DW/BI**

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

■ Entender os leitores:  
■ Identificar suas características demográficas.  
■ Descobrir o que os leitores querem nesse tipo de revista.  
■ Identificar os "melhores" leitores que renovarão suas assinaturas e comprarão produtos dos anunciantes da revista.  
■ Encontrar novos leitores em potencial e torná-los conscientes da revista.  
■ Garantir que a revista atraia os leitores:  
■ Escolher conteúdo interessante e envolvente para a revista.  
■ Tomar decisões de layout e apresentação que maximizem o prazer dos leitores.  
■ Manter altos padrões de escrita e edição enquanto adota um estilo de apresentação consistente.  
■ Monitorar continuamente a precisão dos artigos e as reivindicações dos anunciantes.  
■ Adaptar-se a perfis de leitores em mudança e à disponibilidade de novas contribuições de uma rede de escritores e colaboradores.  
■ Sustentar a publicação:  
■ Atrair anunciantes e gerenciar a revista de forma lucrativa.  
■ Publicar a revista regularmente.  
■ Manter a confiança dos leitores.  
■ Manter os proprietários do negócio felizes.

Você também pode identificar itens que não devem ser objetivos para o editor-chefe da revista, como construir a revista em torno de uma tecnologia de impressão específica ou colocar exclusivamente a energia da gestão em eficiências operacionais, como impor um estilo de escrita técnica que os leitores não entendem facilmente, ou criar um layout intrincado e lotado que é difícil de ler.

Ao construir o negócio editorial sobre uma fundação de atendimento eficaz aos leitores, a revista provavelmente será bem-sucedida. Por outro lado, percorra a lista e imagine o que acontece se você omitir qualquer item; em última análise, a revista teria problemas sérios.

Há paralelos fortes que podem ser traçados entre ser um editor convencional e ser um gerente DW/BI. Impulsionados pelas necessidades do negócio, os gerentes DW/BI devem publicar dados que foram coletados de uma variedade de fontes e editados para qualidade e consistência. A principal responsabilidade é servir os leitores, também conhecidos como usuários de negócios. A metáfora editorial sublinha a necessidade de focar nos seus clientes em vez de apenas nos produtos e processos internos. Embora você use tecnologia para entregar o sistema DW/BI, a tecnologia é, no máximo, um meio para um fim. Como tal, a tecnologia e as técnicas usadas para construir o sistema não devem aparecer diretamente nas suas principais responsabilidades de trabalho.

Agora, reinterprete as responsabilidades do editor de revistas como responsabilidades do gerente DW/BI:

■ Entender os usuários de negócios:  
■ Compreender suas responsabilidades, metas e objetivos no trabalho.  
■ Determinar as decisões que os usuários de negócios querem tomar com a ajuda do sistema DW/BI.  
■ Identificar os "melhores" usuários que tomam decisões eficazes e de alto impacto.  
■ Encontrar novos usuários em potencial e torná-los conscientes das capacidades do sistema DW/BI.  
■ Entregar informações e análises de alta qualidade, relevantes e acessíveis aos usuários de negócios:  
■ Escolher os dados mais robustos e acionáveis para apresentar no sistema DW/BI, cuidadosamente selecionados a partir do vasto universo de possíveis fontes de dados na sua organização.  
■ Tornar as interfaces de usuário e aplicações simples e baseadas em modelos, explicitamente ajustadas aos perfis de processamento cognitivo dos usuários.  
■ Garantir que os dados sejam precisos e confiáveis, rotulando-os de forma consistente em toda a empresa.  
■ Monitorar continuamente a precisão dos dados e das análises.  
■ Adaptar-se a perfis de usuários em mudança, requisitos e prioridades de negócios, juntamente com a disponibilidade de novas fontes de dados.  
■ Sustentar o ambiente DW/BI:  
■ Apropriar-se de parte do crédito pelas decisões de negócios tomadas usando o sistema DW/BI e usar esses sucessos para justificar pessoal e despesas contínuas.  
■ Atualizar o sistema DW/BI regularmente.  
■ Manter a confiança dos usuários de negócios.  
■ Manter os usuários de negócios, patrocinadores executivos e gestão de TI felizes.

Se você fizer um bom trabalho com todas essas responsabilidades, será um ótimo gerente DW/BI! Por outro lado, percorra a lista e imagine o que acontece se você omitir qualquer item. Em última análise, o ambiente teria problemas sérios. Agora contraste essa visão do trabalho de um gerente DW/BI com a descrição do seu próprio trabalho. É provável que a lista anterior esteja mais orientada para questões de usuários e negócios e pode até não parecer um trabalho em TI. Em nossa opinião, isso é o que torna o data warehousing e business intelligence interessante.

**Introdução à Modelagem Dimensional**

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

■ Entregar dados que sejam compreensíveis para os usuários de negócios.  
■ Entregar desempenho rápido em consultas.

A modelagem dimensional é uma técnica antiga para simplificar bancos de dados. Em caso após caso, por mais de cinco décadas, organizações de TI, consultores e usuários de negócios naturalmente gravitaram em direção a uma estrutura dimensional simples para corresponder à necessidade humana fundamental de simplicidade. A simplicidade é crítica porque garante que os usuários possam facilmente entender os dados, assim como permite que o software navegue e entregue resultados rápida e eficientemente.

Imagine uma executiva que descreve seu negócio como: “Vendemos produtos em vários mercados e medimos nosso desempenho ao longo do tempo.” Os projetistas dimensionais ouvem atentamente a ênfase em produto, mercado e tempo. A maioria das pessoas acha intuitivo pensar em um negócio como esse como um cubo de dados, com as bordas rotuladas como produto, mercado e tempo. Imagine fatiar e analisar ao longo de cada uma dessas dimensões. Pontos dentro do cubo são onde as medições, como volume de vendas ou lucro, para aquela combinação de produto, mercado e tempo são armazenadas. A capacidade de visualizar algo tão abstrato como um conjunto de dados de uma maneira concreta e tangível é o segredo da compreensibilidade. Se essa perspectiva parecer simples demais, ótimo! Um modelo de dados que começa simples tem uma chance de permanecer simples no final do projeto. Um modelo que começa complicado certamente será excessivamente complicado no final, resultando em desempenho lento de consultas e rejeição dos usuários de negócios. Albert Einstein capturou a filosofia básica que impulsiona o design dimensional quando disse: “Faça tudo o mais simples possível, mas não mais simples.”

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

A indústria às vezes se refere a modelos 3FN como modelos entidade-relacionamento (ER). Diagramas de entidade-relacionamento (ERDs) são desenhos que comunicam as relações entre tabelas. Tanto os modelos 3FN quanto os modelos dimensionais podem ser representados em ERDs porque ambos consistem em tabelas relacionais unidas; a principal diferença entre 3FN e modelos dimensionais é o grau de normalização. Como ambos os tipos de modelo podem ser apresentados como ERDs, evitamos referir-nos a modelos 3FN como modelos ER; em vez disso, os chamamos de modelos normalizados para minimizar a confusão.

Estruturas normalizadas em 3FN são imensamente úteis no processamento operacional porque uma transação de atualização ou inserção toca o banco de dados em apenas um lugar. No entanto, modelos normalizados são complicados demais para consultas de BI. Os usuários não conseguem entender, navegar ou lembrar de modelos normalizados que se assemelham a um mapa do sistema de rodovias de Los Angeles. Da mesma forma, a maioria dos sistemas de gerenciamento de banco de dados relacionais não pode consultar eficientemente um modelo normalizado; a complexidade das consultas imprevisíveis dos usuários sobrecarrega os otimizadores de banco de dados, resultando em desempenho desastroso de consultas. O uso da modelagem normalizada na área de apresentação do DW/BI prejudica a recuperação intuitiva e de alto desempenho dos dados. Felizmente, a modelagem dimensional resolve o problema de esquemas excessivamente complexos na área de apresentação.

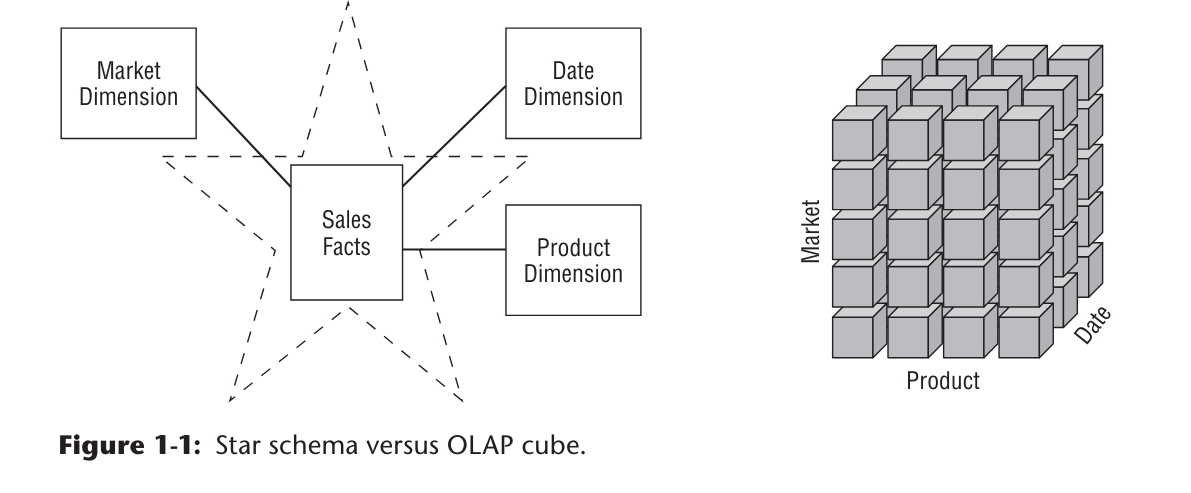
**NOTA:** Um modelo dimensional contém as mesmas informações que um modelo normalizado, mas organiza os dados em um formato que proporciona compreensibilidade do usuário, desempenho em consultas e resiliência a mudanças.

**Esquemas Estrela versus Cubos OLAP**

Modelos dimensionais implementados em sistemas de gerenciamento de banco de dados relacionais são chamados de esquemas estrela por causa de sua semelhança com uma estrutura em forma de estrela. Modelos dimensionais implementados em ambientes de banco de dados multidimensionais são chamados de cubos de processamento analítico online (OLAP), como ilustrado na Figura 1-1.  
Se o seu ambiente DW/BI incluir esquemas estrela ou cubos OLAP, ele aproveita conceitos dimensionais. Tanto estrelas quanto cubos têm um design lógico comum com dimensões reconhecíveis; no entanto, a implementação física difere.

Quando os dados são carregados em um cubo OLAP, eles são armazenados e indexados usando formatos e técnicas projetados para dados dimensionais. Agregações de desempenho ou tabelas de resumo pré-calculadas são frequentemente criadas e gerenciadas pelo mecanismo do cubo OLAP.

Consequentemente, os cubos oferecem desempenho superior em consultas devido aos pré-cálculos, estratégias de indexação e outras otimizações. Os usuários de negócios podem detalhar ou resumir adicionando ou removendo atributos de suas análises com excelente desempenho sem emitir novas consultas. Cubos OLAP também oferecem funções analíticas mais robustas que excedem as disponíveis com SQL. A desvantagem é que você paga um preço de desempenho de carga por essas capacidades, especialmente com grandes conjuntos de dados.

  
 Felizmente, a maioria das recomendações neste livro se aplica independentemente da plataforma de banco de dados relacional ou multidimensional. Embora as capacidades da tecnologia OLAP estejam melhorando continuamente, geralmente recomendamos que informações detalhadas e atômicas sejam carregadas em um esquema estrela; cubos OLAP opcionais são então populados a partir do esquema estrela. Por essa razão, a maioria das técnicas de modelagem dimensional neste livro são descritas em termos de um esquema estrela relacional.

**Considerações sobre Implantação OLAP**

Aqui estão algumas coisas a serem lembradas se você implantar dados em cubos OLAP:

■ Um esquema estrela hospedado em um banco de dados relacional é uma boa base física para a construção de um cubo OLAP, e é geralmente considerado uma base mais estável para backup e recuperação.  
■ Cubos OLAP têm sido tradicionalmente notados por vantagens de desempenho extremo sobre SGBDRs, mas essa distinção se tornou menos importante com os avanços em hardware de computador, como appliances e bancos de dados em memória, e software de SGBDR, como bancos de dados colunares.  
■ As estruturas de dados de cubos OLAP são mais variáveis entre diferentes fornecedores do que SGBDRs, portanto, os detalhes finais de implantação muitas vezes dependem de qual fornecedor de OLAP é escolhido. É tipicamente mais difícil portar aplicações de BI entre diferentes ferramentas de OLAP do que portar aplicações de BI entre diferentes bancos de dados relacionais.  
■ Cubos OLAP tipicamente oferecem opções de segurança mais sofisticadas do que SGBDRs, como limitar o acesso a dados detalhados, mas fornecer acesso mais aberto a dados de resumo.  
■ Cubos OLAP oferecem capacidades de análise significativamente mais ricas do que SGBDRs, que são limitados pelas restrições do SQL. Esta pode ser a principal justificativa para usar um produto OLAP.  
■ Cubos OLAP suportam graciosamente mudanças de tipo 2 de dimensões lentamente mutáveis (que são discutidas no Capítulo 5: Procurement), mas frequentemente precisam ser reprocessados parcial ou totalmente sempre que os dados são sobrescritos usando técnicas alternativas de dimensões lentamente mutáveis.  
■ Cubos OLAP suportam graciosamente tabelas de fatos de transações e instantâneos periódicos, mas não lidam com tabelas de fatos de instantâneos acumulativos devido às limitações na sobrescrita de dados descritas no ponto anterior.  
■ Cubos OLAP tipicamente suportam hierarquias complexas de profundidade indeterminada, como organogramas ou listas de materiais, usando sintaxe de consulta nativa que é superior às abordagens exigidas para SGBDRs.  
■ Cubos OLAP podem impor restrições detalhadas na estrutura das chaves de dimensão que implementam hierarquias de detalhamento em comparação com bancos de dados relacionais.  
■ Alguns produtos OLAP não permitem funções ou aliases dimensionais, exigindo assim que dimensões físicas separadas sejam definidas.

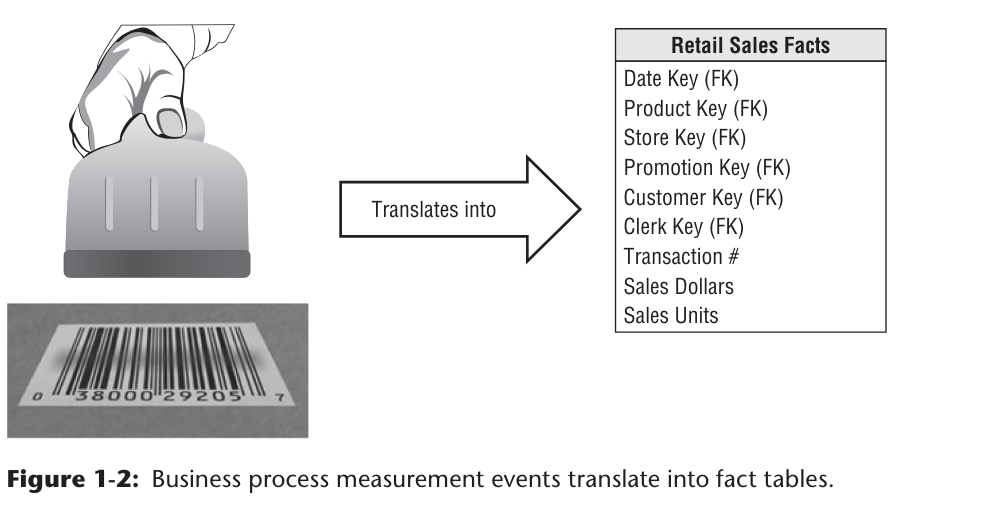
Voltaremos ao mundo da modelagem dimensional em uma plataforma relacional ao considerarmos os dois componentes principais de um esquema estrela.

**Tabelas de Fatos para Medições**

A tabela de fatos em um modelo dimensional armazena as medições 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 único modelo dimensional. Como os dados de medição são esmagadoramente o maior conjunto de dados, eles não devem ser replicados em vários lugares para múltiplas funções organizacionais em toda a empresa. Permitir que usuários de negócios de várias organizações acessem um repositório centralizado único para cada conjunto de dados de medição garante o uso de dados consistentes em toda a empresa.

O termo fato representa uma medida de negócio. Imagine estar no mercado observando produtos sendo vendidos e anotando a quantidade de unidades e o valor das vendas em dólares para cada produto em cada transação de venda. Essas medições são capturadas à medida que os produtos são escaneados no caixa, conforme ilustrado na Figura 1-2.

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

  
**NOTA:** A ideia de que um evento de medição no mundo físico tem uma relação de um-para-um com uma única linha na tabela de fatos correspondente é um princípio fundamental para a modelagem dimensional. Tudo o mais se baseia nessa fundação.

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

A aditividade é crucial porque as aplicações de BI raramente recuperam uma única linha da tabela de fatos. Em vez disso, elas trazem de volta centenas, milhares ou até milhões de linhas de fatos de uma vez, e a coisa mais útil a fazer com tantas linhas é somá-las. Não importa como o usuário fatia 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 até não-aditivos. Fatos semi-aditivos, como saldos de contas, não podem ser somados ao longo da dimensão do 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 bilhões de linhas.

Os fatos são frequentemente descritos como continuamente valorados para ajudar a distinguir o que é um fato versus um atributo de dimensão. O valor das vendas em dólares é continuamente valorado neste exemplo porque pode assumir virtualmente qualquer valor dentro de um amplo intervalo. Como observador, você deve estar no mercado esperando 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 medição textual é uma descrição de algo e é retirada de uma lista discreta de valores. O projetista deve se esforçar para colocar dados textuais em dimensões onde eles podem ser correlacionados mais efetivamente com outros atributos de dimensão textuais e consumir muito menos espaço. Você não deve armazenar informações textuais redundantes em tabelas de fatos. A menos que o texto seja único para cada linha na tabela de fatos, ele pertence à tabela de dimensões. Um fato textual verdadeiro é raro porque o conteúdo imprevisível de um fato textual, como um comentário em texto livre, torna quase impossível analisá-lo.

Referindo-se à tabela de fatos de exemplo na Figura 1-2, se não houver atividade de vendas para um determinado produto, você não coloca 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 esparsidade, 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 de número de linhas, mas estreitas em termos de número de colunas. Dado o seu tamanho, você deve ser criterioso quanto à utilização do espaço da tabela de fatos.

À medida que os exemplos são desenvolvidos ao longo deste livro, você verá que todos os grãos de tabela de fatos se enquadram em uma das três categorias: transação, instantâneo periódico e instantâneo acumulativo. Tabelas de fatos de grão de transação são as mais comuns. Introduziremos tabelas de fatos de transação no Capítulo 3: Vendas no Varejo, e ambos os instantâneos periódicos e acumulativos no Capítulo 4: Inventário.

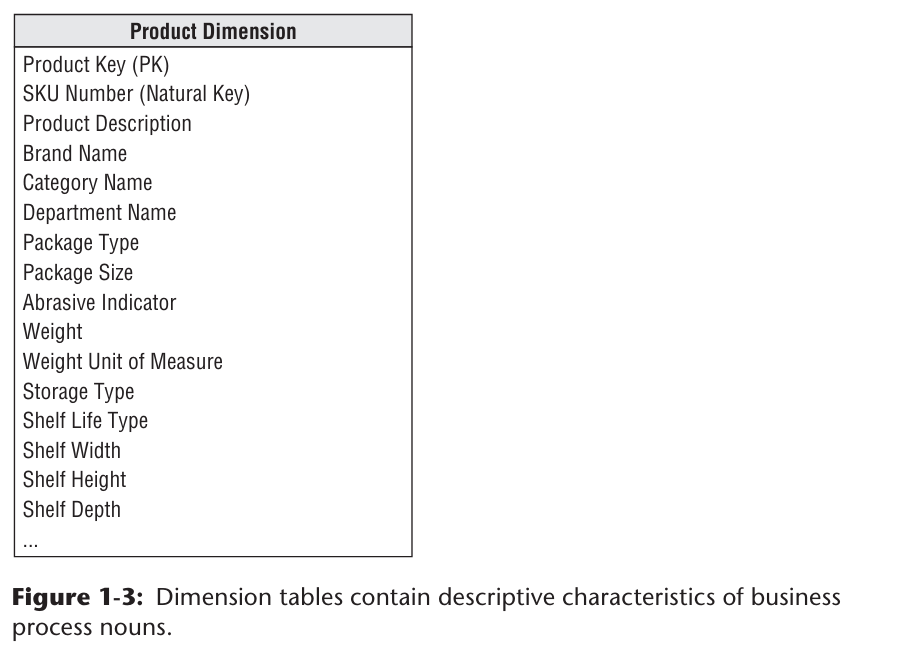
Todas as tabelas de fatos têm duas ou mais chaves estrangeiras (referidas pela notação FK na Figura 1-2) que se conectam às chaves primárias das tabelas de dimensões. Por exemplo, a chave do produto na tabela de fatos sempre corresponde a uma chave de produto específica na tabela de dimensões de produto. Quando todas as chaves na tabela de fatos correspondem corretamente às suas respectivas chaves primárias nas tabelas de dimensões correspondentes, as tabelas satisfazem a integridade referencial. Você acessa a tabela de fatos por meio das tabelas de dimensões unidas a ela.

A tabela de fatos geralmente tem sua própria chave primária composta por um subconjunto das chaves estrangeiras. Esta chave é frequentemente chamada de chave composta. Toda tabela que tem uma chave composta é uma tabela de fatos. Tabelas de fatos expressam relações de muitos para muitos. Todas as outras são tabelas de dimensões.

Normalmente, há um punhado de dimensões que, juntas, identificam de forma única cada linha da tabela de fatos. Após esse subconjunto da lista geral de dimensões ter sido identificado, o restante das dimensões assume um valor único no contexto da chave primária da linha da tabela de fatos. Em outras palavras, elas acompanham.

**Tabelas de Dimensões para Contexto Descritivo**

As tabelas de dimensões são companheiras integrais de uma tabela de fatos. As tabelas de dimensões contêm o contexto textual associado a um evento de medição de um processo de negócios. Elas descrevem o "quem, o quê, onde, quando, como e porquê" associado ao evento.  
Como ilustrado na Figura 1-3, as tabelas de dimensões frequentemente possuem muitas colunas ou atributos. Não é incomum que uma tabela de dimensões tenha de 50 a 100 atributos; embora, algumas tabelas de dimensões naturalmente tenham apenas alguns atributos. As tabelas de dimensões tendem a ter menos linhas do que as tabelas de fatos, mas podem ser largas, com muitas colunas de texto grandes. Cada dimensão é definida por uma única chave primária (referida pela notação PK na Figura 1-3), que serve como base para a integridade referencial com qualquer tabela de fatos à qual esteja unida.



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

Os atributos das tabelas de dimensões desempenham um papel vital no sistema DW/BI. Como são a fonte de praticamente todas as restrições e rótulos de relatórios, os atributos das dimensões são críticos para tornar o sistema DW/BI utilizável e compreensível. Os atributos devem consistir em palavras reais em vez de abreviações criptográficas. Você deve se esforçar para minimizar o uso de códigos nas tabelas de dimensões, substituindo-os por atributos textuais mais verbosos. Você pode já ter treinado os usuários de negócios para memorizar códigos operacionais, mas daqui para frente, minimize sua dependência de notas minúsculas anexadas ao monitor para traduções de códigos.

Você deve disponibilizar decodificações padrão para os códigos operacionais como atributos de dimensões para fornecer rótulos consistentes em consultas, relatórios e aplicações de BI. Os valores de decodificação nunca devem ser enterrados nas aplicações de relatórios, onde a inconsistência é inevitável.

Às vezes, os códigos ou identificadores operacionais têm um significado legítimo para os usuários ou são necessários para se comunicar de volta ao mundo operacional. Nesses casos, os códigos devem aparecer como atributos explícitos de dimensão, além dos descritores textuais amigáveis correspondentes. Os códigos operacionais às vezes têm inteligência embutida neles. Por exemplo, os dois primeiros dígitos podem identificar a linha de negócios, enquanto os dois dígitos seguintes podem identificar a região global. Em vez de forçar os usuários a interrogar ou filtrar subcódigos dentro dos códigos operacionais, extraia os significados embutidos e apresente-os aos usuários como atributos de dimensão separados que podem ser facilmente filtrados, agrupados ou relatados.

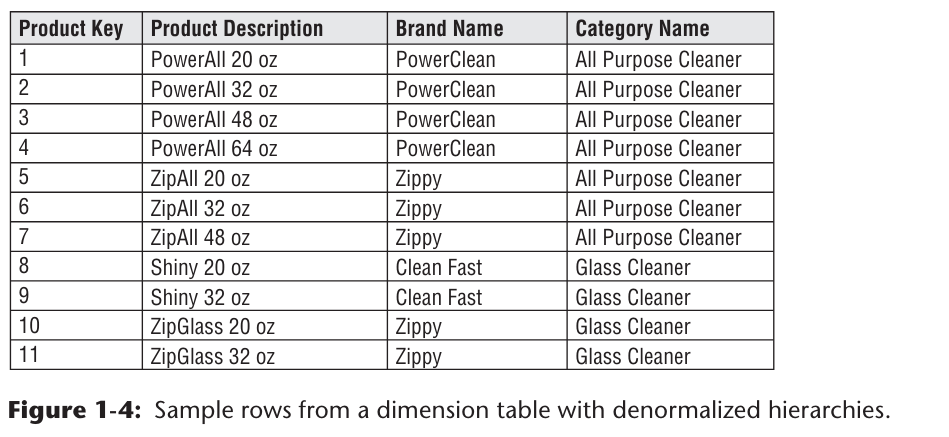
De muitas maneiras, o data warehouse é tão bom quanto os atributos de dimensão; o poder analítico do ambiente DW/BI é diretamente proporcional à qualidade e profundidade dos atributos de dimensão. Quanto mais tempo gasto fornecendo atributos com terminologia de negócios detalhada, melhor. Quanto mais tempo gasto populando os valores de domínio em uma coluna de atributo, melhor. Quanto mais tempo gasto garantindo a qualidade dos valores em uma coluna de atributo, melhor. Atributos de dimensão robustos entregam capacidades analíticas robustas de fatiamento e análise.

**NOTA**: As dimensões fornecem os pontos de entrada para os dados, e os rótulos e agrupamentos finais em todas as análises DW/BI.

Ao realizar a triagem de dados de origem operacionais, às vezes não é claro se um elemento de dados numérico é um fato ou um atributo de dimensão. Você frequentemente toma a decisão perguntando se a coluna é uma medição que assume muitos valores e participa de cálculos (fazendo dela um fato) ou é uma descrição de valor discreto que é mais ou menos constante e participa de restrições e rótulos de linha (fazendo dela um atributo dimensional). Por exemplo, o custo padrão de um produto parece um atributo constante do produto, mas pode ser alterado com tanta frequência que você decide que é mais parecido com um fato medido. Ocasionalmente, você não pode ter certeza da classificação; é possível modelar o elemento de dados de qualquer maneira (ou de ambas as maneiras) como uma prerrogativa do designer.

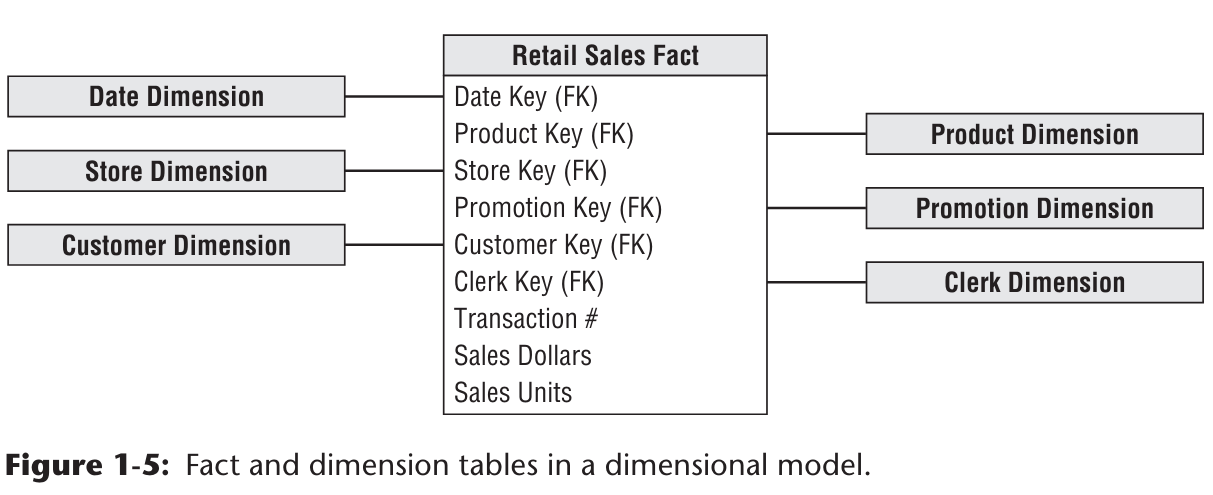
**NOTA**: O dilema do designer sobre se uma quantidade numérica é um fato ou um atributo dimensional raramente é uma decisão difícil. Observações numéricas continuamente valoradas são quase sempre fatos; observações numéricas discretas retiradas de uma pequena lista são quase sempre atributos de dimensão.

A Figura 1-4 mostra que as tabelas de dimensões frequentemente representam relações hierárquicas. Por exemplo, produtos se agrupam em marcas e depois em categorias. Para cada linha na dimensão do produto, você deve armazenar a descrição da marca e da categoria associada. A informação descritiva hierárquica é armazenada de forma redundante no espírito de facilidade de uso e desempenho de consultas. Você deve resistir ao impulso talvez habitual de normalizar dados armazenando apenas o código da marca na dimensão do produto e criando uma tabela de pesquisa de marcas separada, e da mesma forma para a descrição da categoria em uma tabela de pesquisa de categorias separada. Essa normalização é chamada de floco de neve (snowflaking). Em vez de terceira forma normal, as tabelas de dimensões são tipicamente altamente desnormalizadas com relações de muitos para um achatadas dentro de uma única tabela de dimensão. Como as tabelas de dimensões são tipicamente geometricamente menores do que as tabelas de fatos, melhorar a eficiência de armazenamento através da normalização ou floco de neve tem virtualmente nenhum impacto no tamanho total do banco de dados. Você quase sempre deve trocar o espaço da tabela de dimensões por simplicidade e acessibilidade.

  
 Contrariamente ao folclore popular, Ralph Kimball não inventou os termos fato e dimensão. Na medida do possível, a terminologia dimensão e fato originou-se de um projeto de pesquisa conjunto conduzido pela General Mills e a Universidade de Dartmouth na década de 1960. Na década de 1970, tanto a AC Nielsen quanto a IRI usaram consistentemente os termos para descrever suas ofertas de dados sindicados e gravitavam para modelos dimensionais para simplificar a apresentação de suas informações analíticas. Eles entenderam que seus dados não seriam usados a menos que fossem embalados de forma simples. Provavelmente é preciso dizer que nenhuma pessoa inventou a abordagem dimensional. É uma força irresistível no design de bancos de dados que sempre resulta quando o designer coloca a compreensibilidade e o desempenho como os objetivos mais altos.

**Fatos e Dimensões Unidos em um Esquema Estrela**

Agora que você entende tabelas de fatos e dimensões, é hora de unir os blocos de construção em um modelo dimensional, como mostrado na Figura 1-5. Cada processo de negócios é representado por um modelo dimensional que consiste em uma tabela de fatos contendo as medições numéricas do evento, cercada por um halo de tabelas de dimensões que contêm o contexto textual que era verdadeiro no momento em que o evento ocorreu. Esta estrutura em forma de estrela é frequentemente chamada de união estrela (star join), um termo que remonta aos primeiros dias dos bancos de dados relacionais.



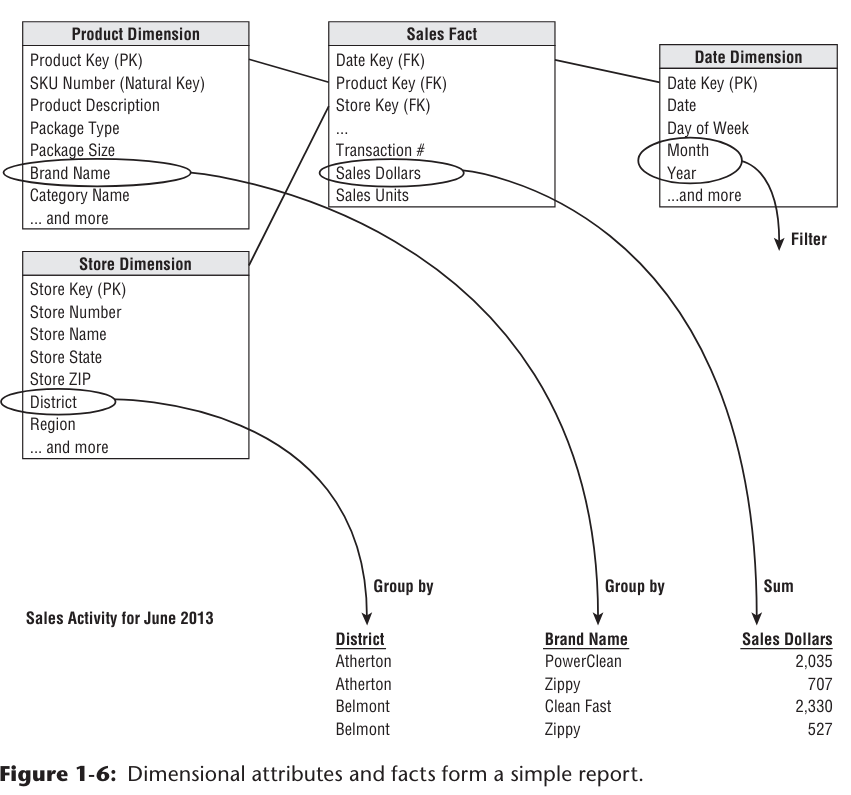
A primeira coisa a notar sobre o esquema dimensional é sua simplicidade e simetria. Obviamente, os usuários de negócios se beneficiam da simplicidade porque os dados são mais fáceis de entender e navegar. O charme do design na Figura 1-5 é que ele é altamente reconhecível para os usuários de negócios.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

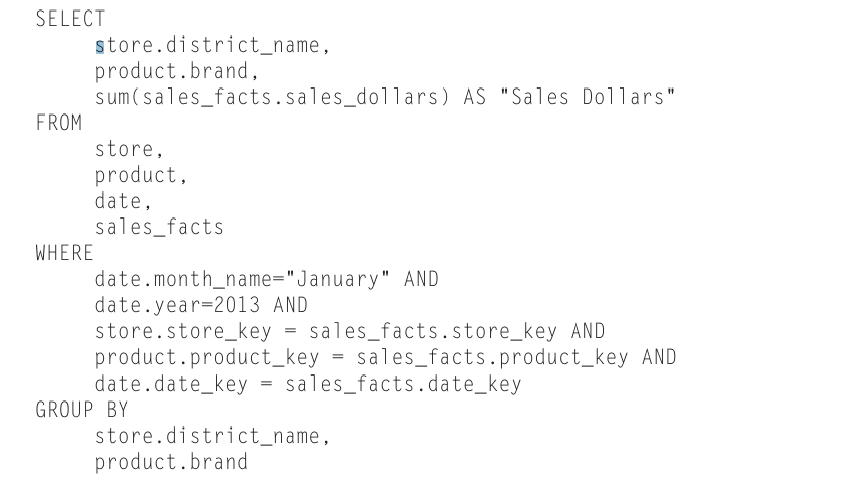
A simplicidade de um modelo dimensional também tem benefícios de desempenho. Os otimizadores de banco de dados processam esses esquemas simples com menos junções de forma mais eficiente. Um mecanismo de banco de dados pode fazer suposições fortes sobre primeiro restringir as tabelas de dimensões fortemente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

Finalmente, os modelos dimensionais são graciosamente extensíveis para acomodar mudanças. A estrutura previsível de um modelo dimensional resiste a 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 tem viés embutido em relação aos padrões de consulta esperados. Não há preferências pelas perguntas de negócios feitas este mês em relação às perguntas feitas no próximo mês. Certamente, você não quer ajustar esquemas se os usuários de negócios sugerirem novas maneiras de analisar seus negócios.

Este livro ilustra repetidamente que os dados mais granulares ou atômicos têm a maior dimensionalidade. Dados atômicos que não foram agregados são os dados mais expressivos; esses dados atômicos devem ser a base para o design de cada tabela de fatos para resistir aos ataques ad hoc dos usuários de negócios em que eles fazem consultas inesperadas. Com modelos dimensionais, você pode adicionar completamente novas dimensões 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, assumindo que o nível de detalhe seja consistente com a tabela de fatos existente. Você pode suplementar tabelas de dimensões preexistentes com novos atributos não previstos. Em cada caso, tabelas existentes podem ser alteradas no lugar simplesmente adicionando novas linhas de dados na tabela ou executando um comando SQL ALTER TABLE. Os dados não precisariam ser recarregados, e as aplicações de BI existentes continuariam a funcionar sem produzir resultados diferentes. Examinamos essa extensibilidade graciosa dos modelos dimensionais mais detalhadamente no Capítulo 3.

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





Se você estudar esse trecho de código linha por linha, as duas primeiras linhas sob a declaração SELECT identificam os atributos de dimensão no relatório, seguidas da 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 filtro do relatório, e o restante declara as junções entre as tabelas de dimensões e de fatos. Finalmente, a cláusula GROUP BY estabelece a agregação dentro do relatório.

**Arquitetura DW/BI de Kimball**

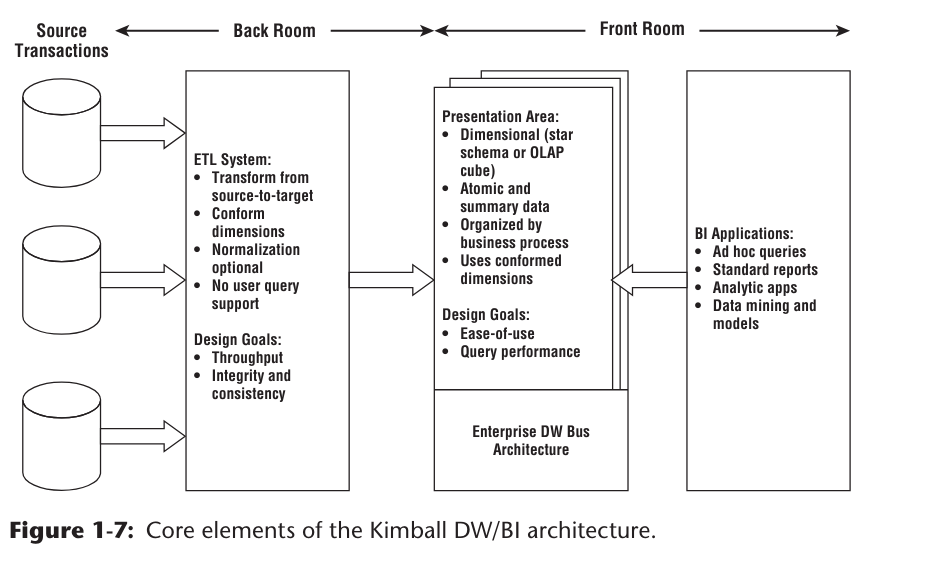
Vamos construir sobre o seu entendimento dos sistemas DW/BI e dos fundamentos da modelagem dimensional investigando os componentes de um ambiente DW/BI baseado na arquitetura de Kimball. Você precisa aprender o significado estratégico de cada componente para evitar confundir seus papéis e funções.

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

### Sistemas de Origem Operacionais

Estes são os sistemas operacionais de registro que capturam as transações do negócio. Pense nos sistemas de origem como estando 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 o desempenho do processamento e a disponibilidade. Consultas operacionais contra sistemas de origem são consultas estreitas, uma por vez, que fazem parte do fluxo normal de transações e são severamente restritas em suas demandas no sistema operacional. É seguro assumir que os sistemas de origem não são consultados de maneira ampla e inesperada como os sistemas DW/BI tipicamente são.

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 aplicações de propósito especial sem qualquer compromisso de compartilhar dados comuns como produto, cliente, geografia ou calendário com outros sistemas operacionais na organização. Claro, um sistema de planejamento de recursos empresariais (ERP) amplamente adotado ou um sistema de gerenciamento de dados mestres operacionais poderia ajudar a resolver essas deficiências.



### Sistema de Extração, Transformação e Carga (ETL)

O sistema de extração, transformação e carga (ETL) do ambiente DW/BI consiste em uma área de trabalho, estruturas de dados instanciadas e um conjunto de processos. O sistema ETL é tudo o que está entre os sistemas de origem operacionais e a área de apresentação DW/BI. Elaboramos a arquitetura dos sistemas ETL e as técnicas associadas no Capítulo 19: Subsistemas e Técnicas ETL, mas queremos apresentar esta peça fundamental do quebra-cabeça do sistema DW/BI.

A extração é o primeiro passo no processo de inserção de dados no ambiente do data warehouse. Extrair significa ler e entender os dados de origem e copiar os dados necessários para o sistema ETL para posterior manipulação. Nesse ponto, os dados pertencem ao data warehouse.

Depois que os dados são extraídos para o sistema ETL, há inúmeras transformações potenciais, como limpar os dados (corrigir erros de ortografia, resolver conflitos de domínio, lidar com elementos ausentes ou parsear em formatos padrão), combinar dados de várias fontes e desduplicar dados. O sistema ETL agrega valor aos dados com essas tarefas de limpeza e conformidade, mudando e melhorando os dados. Além disso, essas atividades podem ser arquitetadas para criar metadados de diagnóstico, levando eventualmente 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 ETL é a estruturação física e carregamento de dados nos modelos dimensionais de destino da área de apresentação. Como a missão principal do sistema ETL é entregar as tabelas de dimensões e fatos na etapa de entrega, esses subsistemas são críticos. Muitos desses subsistemas definidos se concentram no processamento de tabelas de dimensões, como atribuições de chaves substitutas, pesquisas de códigos para fornecer descrições apropriadas, divisão ou combinação de colunas para apresentar os valores de dados apropriados ou unir estruturas de tabelas de terceira forma normal subjacentes em dimensões desnormalizadas e achatadas. Em contraste, as tabelas de fatos são tipicamente grandes e demoradas para carregar, mas prepará-las para a área de apresentação é tipicamente direto. Quando as tabelas de dimensões e fatos em um modelo dimensional foram atualizadas, indexadas, fornecidas com agregados apropriados e com qualidade adicional garantida, a comunidade de negócios é notificada de que os novos dados foram publicados.

Ainda há consternação na indústria sobre se os dados no sistema ETL devem ser reaproveitados em estruturas normalizadas físicas antes de serem carregados nas estruturas dimensionais da área de apresentação para consultas e relatórios. O sistema ETL é tipicamente dominado pelas atividades simples de ordenação e processamento sequencial. Em muitos casos, o sistema ETL não é baseado em tecnologia relacional, mas pode, em vez disso, confiar em um sistema de arquivos planos. Depois de validar os dados para conformidade com as regras de negócios definidas de um-para-um e muitos-para-um, pode ser inútil dar o passo final de construir um banco de dados físico 3NF, apenas antes de transformar os dados novamente em estruturas desnormalizadas para a área de apresentação de BI.

No entanto, há casos em que os dados chegam à porta do sistema ETL em um formato relacional 3NF. Nesses casos, os desenvolvedores do sistema ETL podem se sentir mais confortáveis realizando as tarefas de limpeza e transformação usando estruturas normalizadas. Embora um banco de dados normalizado para processamento 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 depois novamente quando você carrega o modelo dimensional. Obviamente, esse processo em duas etapas requer mais tempo e investimento para o desenvolvimento, mais tempo para o carregamento ou atualização periódica dos dados e mais capacidade para armazenar as múltiplas cópias dos dados. No final das contas, isso geralmente se traduz na necessidade de maiores orçamentos para desenvolvimento, suporte contínuo e plataforma de hardware. Infelizmente, algumas iniciativas DW/BI falharam miseravelmente porque focaram toda a sua energia e recursos na construção das estruturas normalizadas em vez de alocar tempo para desenvolver uma área de apresentação dimensional que suporte a melhoria da tomada de decisões empresariais.

Embora a consistência de dados em toda a empresa seja um objetivo fundamental do ambiente DW/BI, pode haver abordagens eficazes e menos dispendiosas do que criar fisicamente tabelas normalizadas no sistema ETL, se essas estruturas não existirem já.

**NOTA:** É aceitável criar um banco de dados normalizado para suportar os processos ETL; no entanto, este não é o objetivo final. As estruturas normalizadas devem ser proibidas para consultas de usuários porque elas derrotam os objetivos duplos de compreensibilidade e desempenho.

### Área de Apresentação para Suporte ao Business Intelligence

A área de apresentação DW/BI é onde os dados são organizados, armazenados e disponibilizados para consultas diretas por usuários, escritores de relatórios e outras aplicações analíticas de BI.

Como o sistema de back-end ETL é proibido, a área de apresentação é o ambiente DW/BI, na medida em que a comunidade empresarial está preocupada; é tudo o que o negócio vê e toca por meio de suas ferramentas de acesso e aplicações de BI. O título de trabalho preliminar da primeira edição de The Data Warehouse Toolkit era Getting the Data Out. É isso que a área de apresentação com seus modelos dimensionais é. Temos várias opiniões fortes sobre a área de apresentação. Em primeiro lugar, insistimos que os dados sejam apresentados, armazenados e acessados em esquemas dimensionais, seja esquemas estrela relacionais ou cubos OLAP. Felizmente, a indústria amadureceu a ponto de não estarmos mais debatendo essa abordagem; concluiu que a modelagem dimensional é a técnica mais viável para entregar dados aos usuários de DW/BI. Nosso segundo ponto sobre a área de apresentação é que ela deve conter dados detalhados e atômicos. Dados atômicos são necessários para resistir a ataques de consultas ad hoc imprevisíveis dos usuários.

Embora a área de apresentação também possa conter dados agregados para melhorar o desempenho, não é suficiente entregar 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 estão trancados em modelos normalizados. É impraticável esperar que um usuário faça drill-down através dos dados dimensionais quase até o nível mais granular e então perca os benefícios de uma apresentação dimensional no passo final.

Embora os usuários de DW/BI e aplicações possam olhar raramente para um único item de linha em um pedido, eles podem estar muito interessados nos pedidos da semana passada para produtos de um determinado tamanho (ou sabor, tipo de embalagem ou fabricante) para clientes que compraram pela primeira vez nos últimos 6 meses (ou residem em um determinado estado ou têm 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 em constante mudança, você deve fornecer acesso aos detalhes precisos para que eles possam se agregar para abordar as perguntas do momento.

A área de apresentação de dados deve ser estruturada em torno de eventos de medição de processos de negócios. Essa abordagem se alinha naturalmente com os sistemas de captura de dados de origem operacionais. Os modelos dimensionais devem corresponder aos eventos físicos de captura de dados; não devem ser projetados para entregar o relatório do dia. Os processos de negócios de uma empresa cruzam 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 popular 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 conformes comuns. Esta é a base da arquitetura de ônibus de data warehouse empresarial descrita no Capítulo 4. A adesão à arquitetura de ônibus é o último ponto na área de apresentação. Sem dimensões conformes compartilhadas, um modelo dimensional se torna uma aplicação independente. Conjuntos de dados isolados que não podem ser unidos são a praga do movimento DW/BI, pois perpetuam visões incompatíveis da empresa. Se você tiver alguma esperança de construir um ambiente DW/BI robusto e integrado, deve se comprometer com a arquitetura de ônibus empresarial. Quando os modelos dimensionais são projetados com dimensões conformes, eles podem ser prontamente combinados e usados juntos. A área de apresentação em uma solução DW/BI de grande empresa consiste, em última análise, em dezenas de modelos dimensionais com muitas das tabelas de dimensões associadas compartilhadas entre as tabelas de fatos.

Usar a arquitetura de ônibus é o segredo para construir sistemas DW/BI distribuídos. Quando a arquitetura de ônibus é usada como uma estrutura, você pode desenvolver o data warehouse empresarial de forma ágil, descentralizada, com escopo realista e iterativa.

**NOTA:** Os dados na área de apresentação consultável 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 ônibus de data warehouse empresarial. Os dados não devem ser estruturados de acordo com a interpretação dos dados por departamentos individuais.

**Aplicações de Business Intelligence**

O último componente principal da arquitetura DW/BI de Kimball é a aplicação de business intelligence (BI). O termo aplicação de BI refere-se vagamente à gama de capacidades fornecidas aos usuários de negócios para alavancar a área de apresentação para a tomada de decisões analíticas. Por definição, todas as aplicações de BI consultam os dados na área de apresentação DW/BI. Consultar, obviamente, é o ponto principal de usar dados para melhorar a tomada de decisões.  
Uma aplicação de BI pode ser tão simples quanto uma ferramenta de consulta ad hoc ou tão complexa quanto uma aplicação sofisticada de mineração de dados ou modelagem. Ferramentas de consulta ad hoc, por mais poderosas que sejam, podem ser compreendidas e usadas efetivamente por apenas 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 aplicações pré-construídas baseadas em parâmetros e modelos que não exigem que os usuários construam consultas diretamente. Algumas das aplicações mais sofisticadas, como ferramentas de modelagem ou previsão, podem carregar resultados de volta nos sistemas de origem operacionais, no sistema ETL ou na área de apresentação.

**Metáfora do Restaurante para a Arquitetura de Kimball**

Uma de nossas metáforas favoritas reforça a importância de separar o ambiente DW/BI em componentes distintos. Neste caso, vamos considerar 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 à parte. 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 começar a funcionar, uma quantidade significativa de planejamento é necessária para projetar o layout do espaço de trabalho e seus componentes.

A cozinha é organizada com vários objetivos de design em mente. Primeiro, o layout deve ser altamente eficiente. Gerentes de restaurante querem alta produtividade na cozinha. Quando o restaurante está cheio e todos estão com fome, não há tempo para movimentos desperdiçados.

Entregar qualidade consistente da cozinha do restaurante é o segundo objetivo importante. O estabelecimento está condenado se os pratos que saem da cozinha repetidamente não atenderem às expectativas. Para alcançar a consistência, os chefs criam seus molhos especiais uma vez na cozinha, em vez de enviar ingredientes para a mesa onde inevitavelmente ocorrerão variações.

Finalmente, a produção da cozinha, as refeições entregues aos clientes do restaurante, também deve ser de alta integridade. Você não gostaria que alguém pegasse uma intoxicação alimentar ao jantar no seu restaurante. Consequentemente, as cozinhas são projetadas com integridade em mente; a preparação de saladas não ocorre nas mesmas superfícies onde o frango cru é manipulado.  
Assim como qualidade, consistência e integridade são considerações principais ao projetar a cozinha do restaurante, elas também são preocupações contínuas para a gestão diária do restaurante.

Os chefs se esforçam para obter as melhores matérias-primas possíveis. Os produtos adquiridos devem atender aos padrões de qualidade e são rejeitados se não atenderem aos padrões mínimos. A maioria dos restaurantes finos modifica seus menus com base na disponibilidade de ingredientes de qualidade.

O restaurante emprega profissionais habilidosos na sua cozinha, utilizando as ferramentas de seu ofício. Os cozinheiros manipulam facas extremamente afiadas com incrível confiança e facilidade. Eles operam equipamentos poderosos e trabalham ao redor de superfícies extremamente quentes sem incidentes.

Dado o ambiente perigoso, a cozinha dos fundos é proibida aos clientes do restaurante. Coisas acontecem na cozinha que os clientes simplesmente não deveriam ver. Simplesmente não é seguro. Cozinheiros profissionais manipulando facas afiadas não devem ser distraídos pelas perguntas dos clientes. Você também não gostaria que os clientes entrassem na cozinha para mergulhar seus dedos em um molho para ver se querem pedir um prato principal. Para evitar essas intrusões, a maioria dos restaurantes tem uma porta fechada que separa a cozinha da área onde os clientes são servidos.

Mesmo restaurantes que possuem uma cozinha aberta geralmente 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, sempre há salas nos fundos fora de vista onde a preparação menos atraente visualmente ocorre.

O sistema ETL do data warehouse se assemelha à cozinha do restaurante. Os dados de origem são magicamente transformados em informações significativas e apresentáveis. O sistema ETL dos bastidores deve ser planejado e arquitetado muito antes de qualquer dado ser extraído da fonte. Como a cozinha, o sistema ETL é projetado para garantir produtividade. Ele deve transformar dados brutos de origem no modelo de destino de maneira eficiente, minimizando movimentos desnecessários.

Obviamente, o sistema ETL também está altamente preocupado com a qualidade, integridade e consistência dos dados. Os dados de entrada são verificados quanto à qualidade razoável ao entrar. As condições são continuamente monitoradas para garantir que as saídas do ETL tenham alta integridade. Regras de negócios para derivar métricas e atributos de valor agregado de forma consistente são aplicadas uma vez por profissionais habilidosos no sistema ETL, em vez de depender de cada cliente para desenvolvê-las de forma independente. Sim, isso coloca um fardo extra na equipe ETL, mas é feito para entregar um produto melhor e mais consistente aos clientes do DW/BI.

**NOTA:** Um ambiente DW/BI adequadamente projetado troca trabalho nas aplicações de BI da sala da frente em favor do trabalho no sistema ETL dos bastidores. O trabalho na sala da frente deve ser feito repetidamente pelos usuários de negócios, enquanto o trabalho nos bastidores é feito uma vez pela equipe ETL.

Finalmente, o sistema ETL deve ser proibido para os usuários de negócios e desenvolvedores de aplicações de BI. Assim como você não quer que os clientes do restaurante entrem na cozinha e potencialmente consumam comida semi-cozida, você não quer que profissionais ETL ocupados sejam distraídos por perguntas imprevisíveis dos usuários de BI. As consequências podem ser altamente desagradáveis se os usuários mergulharem seus dedos em potes de preparação intermediária enquanto a preparação de dados ainda está em andamento. Como na cozinha do restaurante, atividades ocorrem no sistema ETL que os clientes do DW/BI não deveriam ver. Quando os dados estão prontos e verificados quanto à qualidade para consumo do usuário, são trazidos pela porta para a área de apresentação 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 fatores chave que diferenciam os restaurantes? De acordo com as avaliações e classificações populares de restaurantes, eles geralmente são classificados em quatro qualidades distintas:

■ Comida (qualidade, sabor e apresentação)  
■ Decoração (ambiente atraente e confortável para os clientes)  
■ Serviço (entrega rápida de comida, equipe de suporte atenciosa e comida recebida conforme o pedido)  
■ Custo

A maioria dos clientes foca inicialmente na pontuação da comida ao avaliar opções de jantar. Em primeiro lugar, o restaurante serve boa comida? Esse é o principal produto do restaurante. No entanto, os fatores de decoração, serviço e custo também afetam a experiência geral dos clientes e são considerações ao avaliar se devem ou não comer em um restaurante.

Claro, o principal produto da cozinha do DW/BI são os dados na área de apresentação. Quais dados estão disponíveis? Como no restaurante, o sistema DW/BI fornece "menus" para descrever o que está disponível via metadados, relatórios publicados e aplicações analíticas parametrizadas. Os clientes do DW/BI esperam consistência e alta qualidade. Os dados da área de apresentação devem ser devidamente preparados e seguros para consumo.

A decoração da área de apresentação deve ser organizada para o conforto dos clientes. Deve ser projetada com base nas preferências dos "comensais" de BI, não na equipe de desenvolvimento. O serviço também é crítico no sistema DW/BI. Os dados devem ser entregues, conforme solicitado, prontamente em uma forma que seja atraente para o usuário de negócios ou desenvolvedor de aplicações de BI. Finalmente, o custo é um fator para o sistema DW/BI. A equipe da cozinha pode estar sonhando com refeições elaboradas e caras, mas se não houver mercado a esse preço, o restaurante não sobreviverá.

Se os clientes do restaurante gostarem de sua experiência de jantar, tudo fica perfeito para 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: alto número de clientes, rotatividade de mesas e receita e lucro noturnos, enquanto a rotatividade de funcionários é baixa. As coisas parecem tão boas que o proprietário do restaurante está considerando um local de expansão para lidar com o tráfego. Por outro lado, se os clientes do restaurante não estiverem satisfeitos, 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 recebendo gorjetas. Em pouco tempo, o restaurante fecha.

Os gerentes de restaurante frequentemente verificam proativamente a satisfação dos clientes com a comida e a experiência de jantar. Se um cliente estiver 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 se dar ao luxo de esperar para ouvir reclamações. Muitas vezes, as pessoas abandonarão um restaurante sem sequer expressar suas preocupações. Com o tempo, os gerentes percebem que o número de clientes diminuiu, mas podem nem saber por quê.

Inevitavelmente, os clientes anteriores do DW/BI localizarão outro "restaurante" que melhor atenda às suas necessidades e preferências, desperdiçando os milhões de dólares investidos para projetar, construir e equipar o sistema DW/BI. Claro, você pode evitar esse desfecho infeliz gerenciando o restaurante proativamente; certifique-se de que a cozinha esteja devidamente organizada e utilizada para entregar conforme necessário a comida, decoração, serviço e custo da área de apresentação.

**Arquiteturas Alternativas de DW/BI**

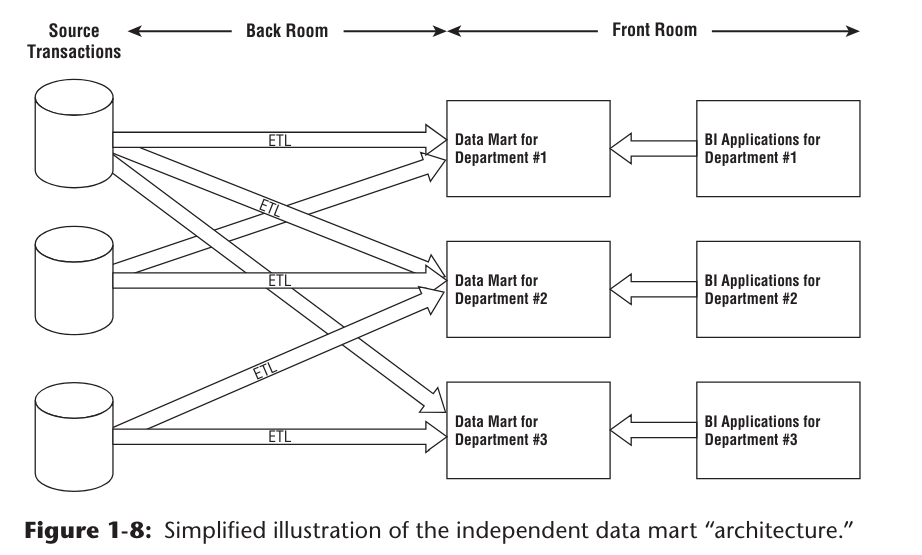
Tendo acabado de descrever a arquitetura de Kimball, vamos discutir várias outras abordagens arquitetônicas de DW/BI. Revisaremos rapidamente as duas principais alternativas à arquitetura de Kimball, destacando as semelhanças e diferenças. Em seguida, fecharemos esta seção focando em uma abordagem híbrida que combina alternativas.

Felizmente, nas últimas décadas, as diferenças entre a arquitetura de Kimball e as alternativas se suavizaram. Ainda mais felizmente, há um papel para a modelagem dimensional, independentemente da sua predisposição arquitetônica.

Reconhecemos que organizações têm construído com sucesso sistemas DW/BI com base nas abordagens defendidas por outros. Acreditamos fortemente que, em vez de encorajar mais consternação sobre nossas diferenças filosóficas, a indústria estaria muito melhor se dedicando energia para garantir que nossos entregáveis de DW/BI sejam amplamente aceitos pelos negócios 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, dados analíticos são implantados em uma base departamental sem preocupação em compartilhar e integrar informações em toda a empresa, conforme ilustrado na Figura 1-8. Tipicamente, 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 de forma isolada, esse data mart departamental atende aos requisitos analíticos do departamento.

Enquanto isso, outro departamento está interessado nos mesmos dados de origem. É extremamente comum que múltiplos departamentos estejam interessados nas mesmas métricas de desempenho resultantes dos eventos de processos de negócios principais de uma organização. Mas, como esse departamento não tem acesso ao data mart inicialmente construído pelo outro departamento, ele segue um caminho semelhante por conta própria, obtendo recursos e construindo 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 surpreendentemente, nenhum dos números combina devido às diferenças nas regras de negócios e na rotulagem. Esses silos analíticos independentes representam uma "arquitetura" DW/BI que é essencialmente não arquitetada.

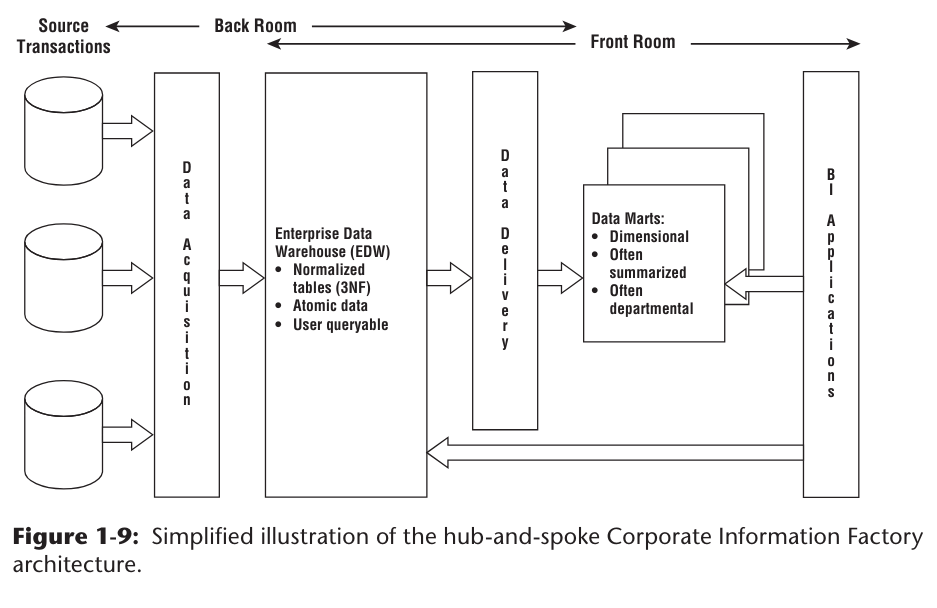
Embora nenhum líder da indústria defenda esses data marts independentes, essa abordagem é prevalente, especialmente em grandes organizações. Ela espelha a maneira como muitas organizações financiam projetos de TI, além de exigir zero governança e coordenação de dados entre organizações. É o caminho de menor resistência para um desenvolvimento rápido a um custo relativamente baixo, pelo menos no curto prazo. Claro, múltiplas extrações não coordenadas das mesmas fontes operacionais e armazenamento redundante de dados analíticos são ineficientes e desperdiçadores a longo prazo. Sem qualquer perspectiva empresarial, essa abordagem independente resulta em várias soluções pontuais isoladas que perpetuam visões incompatíveis do desempenho da organização, resultando em debates e reconciliações organizacionais desnecessárias.

Desencorajamos fortemente a abordagem de data mart independente. No entanto, frequentemente esses data marts independentes adotaram a modelagem dimensional porque estão interessados em entregar dados que sejam fáceis de entender para os negócios e altamente responsivos às consultas.

Portanto, nossos conceitos de modelagem dimensional são frequentemente aplicados nessa arquitetura, apesar do completo desrespeito por alguns dos nossos princípios fundamentais, como focar em detalhes atômicos, construir por processo de negócios em vez de departamento e aproveitar dimensões conformes para consistência e integração empresarial.

**Arquitetura Hub-and-Spoke Corporate Information Factory de Inmon**

A abordagem Hub-and-Spoke Corporate Information Factory (CIF) é defendida por Bill Inmon e outros na indústria. A Figura 1-9 ilustra uma versão simplificada do CIF, focando nos elementos e conceitos principais que merecem discussão.



Com o CIF, os dados são extraídos dos sistemas de origem operacionais e processados através de um sistema ETL, às vezes referido como aquisição de dados. Os dados atômicos resultantes desse processamento são armazenados em um banco de dados 3FN; esse repositório normalizado e atômico é chamado de Data Warehouse Empresarial (EDW) dentro da arquitetura CIF. Embora a arquitetura Kimball permita a normalização opcional para suportar o processamento ETL, o EDW normalizado é uma construção obrigatória no CIF.

Assim como a abordagem Kimball, o CIF defende a coordenação e integração de dados empresariais. O CIF afirma que o EDW normalizado preenche esse papel, enquanto a arquitetura Kimball enfatiza a importância de um ônibus empresarial com dimensões conformes.

**NOTA:** O processo de normalização não trata tecnicamente da integração. A normalização simplesmente cria tabelas físicas que implementam relações de muitos-para-um. A integração, por outro lado, exige que as inconsistências provenientes de fontes separadas sejam resolvidas. Fontes de banco de dados separadas e incompatíveis podem ser normalizadas ao máximo sem abordar a integração. A arquitetura Kimball baseada em dimensões conformes inverte essa lógica e foca na resolução de inconsistências de dados sem exigir explicitamente a normalização.

Organizações que adotaram a abordagem CIF frequentemente têm usuários de negócios acessando o repositório EDW devido ao seu nível de detalhe ou à disponibilidade de dados em tempo hábil. No entanto, processos ETL subsequentes de entrega de dados também populam ambientes de relatórios e análises a jusante para apoiar os usuários de negócios. Embora frequentemente estruturados de forma dimensional, os bancos de dados analíticos resultantes geralmente diferem das estruturas na área de apresentação da arquitetura Kimball em que são frequentemente centrados em departamentos (em vez de organizados em torno de processos de negócios) e populados com dados agregados (em vez de detalhes atômicos). Se os processos ETL de entrega de dados aplicarem regras de negócios além da sumarização básica, como renomeação departamental de colunas ou cálculos alternativos, pode ser difícil conectar esses bancos de dados analíticos ao repositório atômico do EDW.

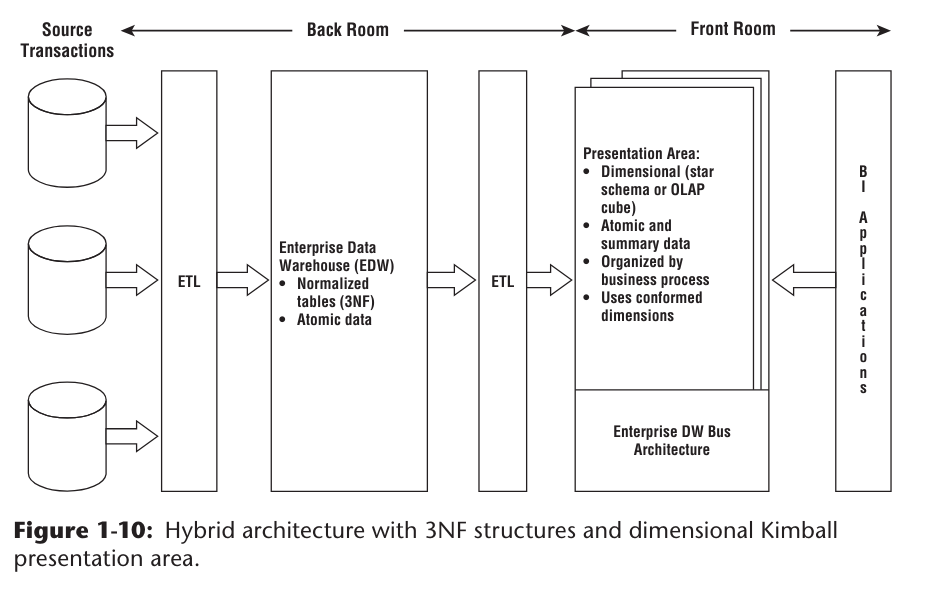
**NOTA**: A forma mais extrema de uma arquitetura CIF pura é inviável como um data warehouse, em nossa opinião. Tal arquitetura bloqueia os dados atômicos em estruturas normalizadas difíceis de consultar, enquanto entrega data marts departamentalmente incompatíveis a diferentes grupos de usuários de negócios. Mas, antes de ficar muito desanimado com essa visão, fique atento à próxima seção.

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

A arquitetura final que merece discussão é a junção das arquiteturas Kimball e Inmon CIF. Como ilustrado na Figura 1-10, essa arquitetura povoa um EDW centrado no CIF que é completamente fora dos limites para usuários de negócios para análise e relatórios. Ele é apenas a fonte para popular uma área de apresentação Kimball-esca na qual os dados são dimensionais, atômicos (complementados por agregados), centrados em processos e conformes à arquitetura de ônibus de data warehouse empresarial.

Alguns defensores dessa abordagem híbrida afirmam que ela é o melhor dos dois mundos. Sim, ela combina as duas abordagens orientadas a empresas. Pode alavancar um investimento pré-existente em um repositório integrado, enquanto aborda os problemas de desempenho e usabilidade associados ao EDW 3FN descarregando consultas para a área de apresentação dimensional. E, como o resultado final para os usuários de negócios e aplicações de BI é construído com base nos princípios de Kimball, quem pode argumentar contra a abordagem?

Se você já investiu na criação de um EDW 3FN, mas ele 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 está começando do zero, a abordagem híbrida provavelmente custará mais tempo e dinheiro, tanto durante o desenvolvimento quanto na operação contínua, dada a movimentação múltipla de dados e o armazenamento redundante de detalhes atômicos. Se você tem 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.



**Mitos da Modelagem Dimensional**

Apesar da ampla aceitação da modelagem dimensional, algumas percepções equivocadas persistem na indústria. Essas falsas afirmações são uma distração, especialmente quando você deseja alinhar sua equipe em torno de práticas recomendadas comuns. Se as pessoas em sua organização continuamente lançarem críticas sobre modelagem dimensional, esta seção deve estar na lista de leitura recomendada; suas percepções podem estar obscurecidas por esses equívocos comuns.  
  
**Mito 1: Modelos Dimensionais são Apenas para Dados Resumidos**

Esse primeiro mito é frequentemente a causa raiz de modelos dimensionais mal projetados. Como você não pode prever todas as perguntas feitas pelos usuários de negócios, você precisa fornecer a eles acesso consultável aos dados mais detalhados para que possam agregá-los com base na pergunta de negócios. Dados no nível mais baixo de detalhe são praticamente imunes a surpresas ou mudanças. Dados resumidos devem complementar os detalhes granulares apenas para proporcionar um desempenho melhor para consultas comuns, mas não substituir os detalhes.

Um corolário relacionado a esse primeiro mito é que apenas uma quantidade limitada de dados históricos deve ser armazenada em estruturas dimensionais. Nada em um modelo dimensional proíbe o armazenamento de um histórico substancial. A quantidade de histórico disponível em modelos dimensionais deve ser impulsionada apenas pelos requisitos do negócio.

**Mito 2: Modelos Dimensionais são Departamentais, Não Empresariais**

Em vez de traçar limites com base em departamentos organizacionais, os modelos dimensionais devem ser organizados em torno de processos de negócios, como pedidos, faturas e chamadas de serviço. Múltiplas funções de negócios frequentemente querem analisar as mesmas métricas resultantes de um único processo de negócios. Extrações múltiplas dos mesmos dados de origem que criam múltiplos bancos de dados analíticos inconsistentes devem ser evitadas.

**Mito 3: Modelos Dimensionais Não São Escaláveis**

Modelos dimensionais são extremamente escaláveis. Tabelas de fatos frequentemente têm bilhões de linhas; tabelas de fatos contendo 2 trilhões de linhas foram reportadas. Os fornecedores de bancos de dados abraçaram completamente o DW/BI e continuam a incorporar capacidades em seus produtos para otimizar a escalabilidade e o desempenho dos modelos dimensionais. Tanto modelos normalizados quanto dimensionais contêm as mesmas informações e relações de dados; o conteúdo lógico é idêntico. Cada relação de dados expressa em um modelo pode ser expressa com precisão no outro. Tanto modelos normalizados quanto dimensionais podem responder exatamente às mesmas perguntas, embora com diferentes níveis de dificuldade.

**Mito 4: Modelos Dimensionais são Apenas para Uso Previsível**

Modelos dimensionais não devem ser projetados focando em relatórios ou análises predefinidos; o design deve se concentrar em processos de medição. Obviamente, é importante considerar os requisitos de filtragem e rotulagem da aplicação de BI. Mas você não deve projetar uma lista dos dez melhores relatórios em um vácuo porque essa lista está destinada a mudar, tornando o modelo dimensional um alvo móvel. A chave é focar nos eventos de medição dos processos de negócios da organização que são tipicamente estáveis, ao contrário das análises que estão em constante evolução.

Um corolário relacionado é que os modelos dimensionais não respondem às mudanças nas necessidades dos negócios. Pelo contrário, devido à sua simetria, as estruturas dimensionais são extremamente flexíveis e adaptáveis a mudanças. O segredo para a flexibilidade das consultas é construir tabelas de fatos no nível mais granular. Modelos dimensionais que fornecem apenas dados resumidos estão destinados a serem problemáticos; os usuários encontram barreiras analíticas quando tentam aprofundar-se nos detalhes que não estão disponíveis nas tabelas resumidas. Os desenvolvedores também encontram barreiras porque não podem facilmente acomodar novas dimensões, atributos ou fatos com essas tabelas resumidas prematuramente. O ponto de partida correto para seus modelos dimensionais é expressar dados no menor detalhe possível para máxima flexibilidade e extensibilidade. Lembre-se, quando você pressupõe a pergunta de negócios, provavelmente pré-resume os dados, o que pode ser fatal a longo prazo. Como o arquiteto Mies van der Rohe é creditado por dizer, "Deus está nos detalhes". Entregar modelos dimensionais populados com os dados mais detalhados possíveis garante máxima flexibilidade e extensibilidade. Entregar qualquer coisa menos nos seus modelos dimensionais mina a fundação necessária para uma inteligência de negócios robusta.

**Mito 5: Modelos Dimensionais Não Podem Ser Integrados**

Modelos dimensionais certamente podem ser integrados se eles conformarem à arquitetura de ônibus de data warehouse empresarial. Dimensões conformes são construídas e mantidas como dados mestres centralizados e persistentes no sistema ETL e depois reutilizados em modelos dimensionais para permitir a integração de dados e garantir consistência semântica. A integração de dados depende de rótulos, valores e definições padronizadas. É um trabalho árduo alcançar consenso organizacional e depois implementar as regras ETL correspondentes, mas você não pode evitar o esforço, independentemente de estar populando modelos normalizados ou dimensionais.  
 Bancos de dados na área de apresentação que não aderem à arquitetura de ônibus com dimensões conformes levam a soluções isoladas. 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 foca na modelagem dimensional para projetar bancos de dados na área de apresentação 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 DW/BI, você precisa ouvir e depois sintetizar as descobertas em torno dos processos de negócios. Às vezes, as equipes se deixam levar ao focar em um conjunto de relatórios ou indicadores de painel necessários. Em vez disso, você deve constantemente se perguntar sobre os eventos de medição dos processos de negócios que produzem os relatórios ou métricas do painel. Ao especificar o escopo do projeto, você deve se manter firme para focar em um único processo de negócios por projeto e não se comprometer a implementar um painel que abranja vários deles em uma única iteração.

Embora seja crucial que a equipe DW/BI se concentre nos processos de negócios, é igualmente importante alinhar a TI e a gestão de negócios na mesma linha de pensamento. Devido às políticas históricas de financiamento de TI, o negócio pode estar mais familiarizado com implantações de dados departamentais. Você precisa mudar a mentalidade deles sobre a implementação do DW/BI para uma perspectiva de processo. Ao priorizar oportunidades e desenvolver o roadmap do DW/BI, os processos de negócios são a unidade de trabalho. Felizmente, a gestão de negócios tipicamente abraça essa abordagem porque reflete seu pensamento sobre indicadores-chave de desempenho.

Além disso, eles viveram com as inconsistências, debates incessantes e reconciliações intermináveis causadas pela abordagem departamental, então estão prontos para uma nova tática. Trabalhando com parceiros de liderança de negócios, classifique cada processo de negócios com base em valor de negócios e viabilidade, depois aborde os processos com as maiores pontuações de impacto e viabilidade primeiro. Embora a priorização seja uma atividade conjunta com o negócio, sua compreensão subjacente dos processos de negócios da organização é essencial para sua eficácia e ação subsequente.

Se for encarregado de redigir a arquitetura de dados do sistema DW/BI, você precisa se familiarizar com os processos da organização, junto com os dados de dimensão descritivos mestres associados. O principal entregável para essa atividade, a matriz de ônibus de data warehouse empresarial, será totalmente avaliada no Capítulo 4. A matriz também serve como uma ferramenta útil para promover os benefícios potenciais de uma plataforma de gerenciamento de dados mestres mais rigorosa.

Programas de governança ou administração de dados devem focar primeiro nas principais dimensões. Dependendo da indústria, a lista pode incluir data, cliente, produto, funcionário, instalação, fornecedor, aluno, corpo docente, conta e assim por diante. Pensar sobre os substantivos centrais usados para descrever o negócio se traduz em uma lista de esforços de governança de dados a serem liderados por especialistas em assuntos da comunidade empresarial. Estabelecer responsabilidades de governança de dados para esses substantivos é a chave para eventualmente implantar dimensões que entreguem consistência e atendam às necessidades dos negócios para filtragem, agrupamento e rotulagem analítica. Dimensões robustas se traduzem em sistemas DW/BI robustos.

Como você pode ver, a motivação fundamental para a modelagem dimensional está em primeiro plano muito antes de você projetar esquemas estrela ou cubos OLAP. Da mesma forma, o modelo dimensional permanecerá em destaque durante os projetos subsequentes do sistema ETL e das aplicações de BI. Os conceitos de modelagem dimensional ligam as comunidades de negócios e técnicas juntas enquanto projetam conjuntamente os entregáveis DW/BI. Vamos elaborar essas ideias no Capítulo 17: Visão Geral do Ciclo de Vida DW/BI de Kimball e no Capítulo 18: Processo e Tarefas de Modelagem Dimensional, mas queríamos plantar as sementes cedo para que tenham tempo de germinar.

**Considerações Ágeis**

Atualmente, há um interesse significativo na indústria DW/BI em práticas de desenvolvimento ágeis. Correndo o risco de simplificação excessiva, as metodologias ágeis focam em incrementos de trabalho de tamanho gerenciável 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 entregáveis prometidos em meses ou anos. Parece bom, não é?

Muitos dos princípios fundamentais das metodologias ágeis se alinham com as melhores práticas de Kimball, incluindo:

■ Foco na entrega de valor de negócios. Este tem sido o mantra de Kimball por décadas.  
■ Valorização da colaboração entre a equipe de desenvolvimento e as partes interessadas do negócio. Como o campo ágil, encorajamos fortemente uma parceria estreita com os negócios.  
■ Ênfase na comunicação contínua face a face, feedback e priorização com as partes interessadas do negócio.  
■ Adaptação rápida às necessidades inevitavelmente evolutivas.  
■ Abordagem do desenvolvimento de maneira iterativa e incremental.

Embora essa lista seja convincente, uma crítica comum às abordagens ágeis é a falta de planejamento e arquitetura, juntamente com desafios contínuos de governança. A matriz de ônibus de data warehouse empresarial é uma ferramenta poderosa para abordar essas deficiências. A matriz de ônibus fornece uma estrutura e um plano mestre para o desenvolvimento ágil, além de identificar as dimensões descritivas comuns reutilizáveis que proporcionam tanto consistência de dados quanto entrega mais rápida ao mercado. Com a mistura colaborativa correta de partes interessadas de negócios e TI em uma sala, a matriz de ônibus de data warehouse empresarial pode ser produzida em um tempo relativamente curto. O trabalho de desenvolvimento incremental pode produzir componentes da estrutura até que funcionalidades suficientes estejam disponíveis e, em seguida, sejam liberadas para a comunidade empresarial.

Alguns clientes e alunos lamentam que, embora queiram entregar dimensões conformes definidas de forma consistente em seus ambientes DW/BI, "simplesmente não é viável". Eles explicam que fariam isso se pudessem, mas com o foco em técnicas de desenvolvimento ágil, é "impossível" dedicar tempo para obter consenso organizacional sobre dimensões conformes. Argumentamos que as dimensões conformes permitem o desenvolvimento ágil de DW/BI, juntamente com a tomada de decisões ágil. À medida que você desenvolve o portfólio de dimensões conformes mestres, a engrenagem de desenvolvimento começa a girar mais rápido e mais rápido. O tempo de entrega ao mercado para uma nova fonte de dados de processo de negócios diminui à medida que os desenvolvedores reutilizam as dimensões conformes existentes. Em última análise, o novo desenvolvimento de ETL foca quase exclusivamente na entrega de mais tabelas de fatos porque as tabelas de dimensões associadas já estão prontas para uso.

Sem uma estrutura como a matriz de ônibus de data warehouse empresarial, algumas equipes DW/BI caíram na armadilha de usar técnicas ágeis para criar soluções analíticas ou de relatórios em um vácuo. Na maioria das situações, a equipe trabalhou com um pequeno conjunto de usuários para extrair um conjunto limitado de dados de origem e torná-los disponíveis para resolver seus problemas únicos. O resultado é frequentemente um conjunto de dados isolado que outros não podem aproveitar, ou pior ainda, entrega dados que não se relacionam com as outras informações analíticas da organização. Encorajamos a agilidade, quando apropriada, no entanto, a construção de conjuntos de dados isolados deve ser evitada. Como na maioria das coisas na vida, moderação e equilíbrio entre os extremos é quase sempre prudente.

**Resumo**

Neste capítulo, discutimos os objetivos predominantes para os sistemas DW/BI e os conceitos fundamentais da modelagem dimensional. Comparamos a arquitetura DW/BI de Kimball com várias alternativas. Encerramos o capítulo identificando equívocos comuns que alguns ainda mantêm sobre a modelagem dimensional, apesar de sua ampla aceitação na indústria, e desafiamos você a pensar dimensionalmente além da modelagem de dados. No próximo capítulo, você terá uma visão turbinada dos padrões e técnicas de modelagem dimensional e começará a colocar esses conceitos em ação no seu primeiro estudo de caso no Capítulo 3.

**Capítulo 2.**

**Visão Geral das Técnicas de Modelagem Dimensional de Kimball**

Começando com a primeira edição de The Data Warehouse Toolkit (Wiley, 1996), o Grupo Kimball definiu o conjunto completo de técnicas para modelar dados de forma dimensional. Nas duas primeiras edições deste livro, sentimos que as técnicas precisavam ser apresentadas através de casos de uso familiares de várias indústrias. Embora ainda consideremos os casos de uso de negócios uma abordagem pedagógica essencial, as técnicas se tornaram tão padronizadas que alguns modeladores dimensionais invertem a lógica começando com a técnica e, em seguida, prosseguindo para o caso de uso para contexto. Tudo isso é uma boa notícia!

As técnicas de Kimball foram aceitas como melhores práticas da indústria. Como evidência, alguns ex-alunos da Kimball University publicaram seus próprios livros de modelagem dimensional. Esses livros geralmente explicam as técnicas de Kimball com precisão, mas é um sinal da resiliência de nossas técnicas que livros alternativos não estenderam significativamente a biblioteca de técnicas nem ofereceram orientações conflitantes.

Este capítulo é a lista "oficial" das Técnicas de Modelagem Dimensional de Kimball dos inventores desses padrões de design. Não esperamos que você leia este capítulo do início ao fim inicialmente. Mas pretendemos que o capítulo seja uma referência para nossas técnicas. Com cada técnica, incluímos referências a capítulos subsequentes para explicações adicionais e ilustrações baseadas nos casos de uso motivadores.

**Conceitos Fundamentais**

As técnicas nesta seção devem ser consideradas durante todos os designs dimensionais. Quase todos os capítulos do livro fazem referência ou ilustram os conceitos desta seção.

**Coleta de Requisitos de Negócios e Realidades dos Dados**

Antes de iniciar um esforço de modelagem dimensional, a equipe precisa entender as necessidades do negócio, bem como as realidades dos dados de origem subjacentes. Você descobre os requisitos através de sessões com representantes de negócios para entender seus objetivos com base em indicadores-chave de desempenho, questões de negócios convincentes, processos de tomada de decisão e necessidades analíticas de suporte. Ao mesmo tempo, as realidades dos dados são descobertas por meio de reuniões com especialistas em sistemas de origem e fazendo uma análise de dados em alto nível para avaliar a viabilidade dos dados.

**Workshops Colaborativos de Modelagem Dimensional**

Os modelos dimensionais devem ser projetados em colaboração com especialistas no assunto e representantes de governança de dados do negócio. O modelador de dados está no comando, mas o modelo deve ser desenvolvido através de uma série de workshops altamente interativos com representantes de negócios. Esses workshops fornecem outra oportunidade para detalhar os requisitos com o negócio. Modelos dimensionais não devem ser projetados isoladamente por pessoas que não entendem completamente o negócio e suas necessidades; a colaboração é crítica!  
  
**Processo de Design Dimensional em Quatro Etapas**

As quatro principais decisões tomadas durante o design de um modelo dimensional incluem:

* Selecionar o processo de negócios.
* Declarar o grão.
* Identificar as dimensões.
* Identificar os fatos.

As respostas a essas perguntas são determinadas considerando as necessidades do negócio juntamente com as realidades dos dados de origem subjacentes durante as sessões de modelagem colaborativa. Após as declarações do processo de negócios, grão, dimensão e fato, a equipe de design determina os nomes das tabelas e colunas, valores de domínio de amostra e regras de negócios. Representantes de governança de dados de negócios devem participar dessa atividade de design detalhado para garantir a aceitação dos negócios.

**Processos de Negócios**

Os processos de negócios são as atividades operacionais realizadas pela sua organização, como fazer um pedido, processar um sinistro de seguro, registrar alunos para uma aula ou tirar um instantâneo de cada conta a cada mês. Eventos de processos de negócios geram ou capturam métricas de desempenho que se traduzem em fatos em uma tabela de fatos. A maioria das tabelas de fatos foca nos resultados de um único processo de negócios. Escolher o processo é importante porque define um alvo de design específico e permite que o grão, dimensões e fatos sejam declarados. Cada processo de negócios corresponde a uma linha na matriz de ônibus de data warehouse empresarial.

**Grão**

Declarar o grão é o passo crucial em um design dimensional. O grão estabelece exatamente o que uma única linha da tabela de fatos representa. A declaração do grão se torna um contrato vinculativo no design. O grão deve ser declarado antes de escolher dimensões ou fatos porque cada dimensão ou fato candidato deve ser consistente com o grão. Essa consistência impõe uma uniformidade em todos os designs dimensionais que é crítica para o desempenho e facilidade de uso das aplicações de BI. Grão atômico refere-se ao nível mais baixo no qual os dados são capturados por um determinado processo de negócios. Incentivamos fortemente que você comece focando em dados com grão atômico porque eles resistem ao ataque de consultas de usuários imprevisíveis; grãos resumidos são importantes para ajuste de desempenho, mas pressupõem as perguntas comuns do negócio. Cada grão de tabela de fatos proposto resulta em uma tabela física separada; grãos diferentes não devem ser misturados na mesma tabela de fatos.

**Dimensões para Contexto Descritivo**

As dimensões fornecem o contexto de “quem, o que, onde, quando, por que e como” ao redor de um evento de processo de negócios. As tabelas de dimensões contêm os atributos descritivos usados por aplicações de BI para filtrar e agrupar os fatos. Com o grão de uma tabela de fatos firmemente em mente, todas as possíveis dimensões podem ser identificadas. Sempre que possível, uma dimensão deve ter um único valor quando associada a uma determinada linha de fato. Tabelas de dimensões são às vezes chamadas de "alma" do data warehouse porque contêm os pontos de entrada e rótulos descritivos que permitem que o sistema DW/BI seja aproveitado para análise de negócios. Um esforço desproporcional é dedicado à governança de dados e ao desenvolvimento de tabelas de dimensões porque elas são os motores da experiência do usuário de BI.  
  
**Fatos para Medições**

Os fatos são as medições resultantes de um evento de processo de negócios e são quase sempre numéricos. Uma única linha de tabela de fatos tem uma relação um-para-um com um evento de medição, conforme descrito pelo grão da tabela de fatos. Assim, uma tabela de fatos corresponde a um evento observável físico, e não às demandas de um relatório específico. Dentro de uma tabela de fatos, apenas os fatos consistentes com o grão declarado são permitidos. Por exemplo, em uma transação de vendas no varejo, a quantidade de um produto vendido e seu preço estendido são bons fatos, enquanto o salário do gerente da loja não é permitido.

**Esquemas Estrela e Cubos OLAP**

Esquemas estrela são estruturas dimensionais implantadas em um sistema de gerenciamento de banco de dados relacional (RDBMS). Eles caracteristicamente consistem em tabelas de fatos ligadas a tabelas de dimensões associadas através de relações chave primária/estrangeira. Um cubo de processamento analítico online (OLAP) é uma estrutura dimensional implementada em um banco de dados multidimensional; ele pode ser equivalente em conteúdo ou, mais frequentemente, derivado de um esquema estrela relacional. Um cubo OLAP contém atributos dimensionais e fatos, mas é acessado através de linguagens com mais capacidades analíticas do que SQL, como XMLA e MDX. Cubos OLAP estão incluídos nesta lista de técnicas básicas porque um cubo OLAP é frequentemente o passo final na implantação de um sistema dimensional DW/BI ou pode existir como uma estrutura agregada baseada em um esquema estrela relacional mais atômico.  
  
**Extensões Graciosas para Modelos Dimensionais**

Modelos dimensionais são resilientes quando as relações de dados mudam. Todas as seguintes mudanças podem ser implementadas sem alterar qualquer consulta ou aplicação de BI existente e sem qualquer mudança nos resultados das consultas.

Fatos consistentes com o grão de uma tabela de fatos existente podem ser adicionados criando novas colunas.

Dimensões podem ser adicionadas a uma tabela de fatos existente criando novas colunas de chave estrangeira, presumindo que não alterem o grão da tabela de fatos.  
Atributos podem ser adicionados a uma tabela de dimensões existente criando novas colunas.

O grão de uma tabela de fatos pode ser tornado mais atômico adicionando atributos a uma tabela de dimensões existente e, em seguida, restando a tabela de fatos no grão inferior, tomando cuidado para preservar os nomes das colunas existentes nas tabelas de fatos e dimensões.

**Técnicas Básicas de Tabelas de Fatos**

As técnicas nesta seção se aplicam a todas as tabelas de fatos. Existem ilustrações de tabelas de fatos em quase todos os capítulos.

**Estrutura da Tabela de Fatos**

Uma tabela de fatos contém as medidas numéricas produzidas por um evento de medição operacional no mundo real. No nível mais baixo de grão, uma linha de tabela de fatos corresponde a um evento de medição e vice-versa. Assim, o design fundamental de uma tabela de fatos é inteiramente baseado em uma atividade física e não é influenciado pelos relatórios que eventualmente podem ser produzidos. Além de medidas numéricas, uma tabela de fatos sempre contém chaves estrangeiras para cada uma de suas dimensões associadas, bem como chaves de dimensões degeneradas opcionais e carimbos de data/hora. Tabelas de fatos são o alvo primário de cálculos e agregações dinâmicas resultantes de consultas.

**Fatos Aditivos, Semi-Aditivos e Não Aditivos**

As medidas numéricas em uma tabela de fatos caem em três categorias. Os fatos mais flexíveis e úteis são totalmente aditivos; medidas aditivas podem ser somadas em qualquer uma das dimensões associadas à tabela de fatos. Medidas semi-aditivas podem ser somadas em algumas dimensões, mas não em todas; valores de saldo são fatos semi-aditivos comuns porque são aditivos em todas as dimensões, exceto no tempo. Finalmente, algumas medidas são completamente não aditivas, como razões. Uma boa abordagem para fatos não aditivos é, onde possível, armazenar os componentes totalmente aditivos da medida não aditiva e somar esses componentes no conjunto de respostas final antes de calcular o fato não aditivo final. Esse cálculo final é frequentemente feito na camada de BI ou cubo OLAP.

**Nulos em Tabelas de Fatos**

Medidas com valor nulo comportam-se graciosamente em tabelas de fatos. As funções de agregação (SUM, COUNT, MIN, MAX e AVG) todas fazem o "certo" com fatos nulos. No entanto, nulos devem ser evitados nas chaves estrangeiras da tabela de fatos porque esses nulos causariam automaticamente uma violação de integridade referencial. Em vez de uma chave estrangeira nula, a tabela de dimensões associada deve ter uma linha padrão (e chave substituta) representando a condição desconhecida ou não aplicável.

**Fatos Conformes**

Se a mesma medida aparecer em tabelas de fatos separadas, deve-se tomar cuidado para garantir que as definições técnicas dos fatos sejam idênticas se forem ser comparadas ou calculadas juntas. Se as definições de fatos separadas forem consistentes, os fatos conformes devem ser nomeados de forma idêntica; mas se forem incompatíveis, devem ser nomeados de forma diferente para alertar os usuários de negócios e as aplicações de BI.

**Tabelas de Fatos de Transação**

Uma linha em uma tabela de fatos de transação corresponde a um evento de medição em um ponto no espaço e no tempo. Tabelas de fatos de transação com grão atômico são as tabelas de fatos mais dimensionais e expressivas; essa robusta dimensionalidade permite o máximo de fatiação e dicing dos dados de transação. Tabelas de fatos de transação podem ser densas ou esparsas porque linhas existem apenas se as medições ocorrerem. Essas tabelas de fatos sempre contêm uma chave estrangeira para cada dimensão associada e, opcionalmente, contêm carimbos de tempo precisos e chaves de dimensões degeneradas. As medidas numéricas devem ser consistentes com o grão de transação.  
  
**Tabelas de Fatos de Snapshot Periódico**

Uma linha em uma tabela de fatos de snapshot periódico resume muitos eventos de medição ocorrendo em um período padrão, como um dia, uma semana ou um mês. O grão é o período, não a transação individual. Tabelas de fatos de snapshot periódicas frequentemente contêm muitos fatos porque qualquer evento de medição consistente com o grão da tabela de fatos é permitido. Essas tabelas de fatos são uniformemente densas em suas chaves estrangeiras porque mesmo que nenhuma atividade ocorra durante o período, uma linha é tipicamente inserida na tabela de fatos contendo um zero ou nulo para cada fato.

**Tabelas de Fatos de Snapshot Acumulativo**

Uma linha em uma tabela de fatos de snapshot acumulativo resume os eventos de medição ocorrendo em etapas previsíveis entre o início e o fim de um processo. Processos de pipeline ou fluxo de trabalho, como o cumprimento de pedidos ou processamento de sinistros, que têm um ponto de início definido, etapas intermediárias padrão e ponto final definido podem ser modelados com esse tipo de tabela de fatos. Há uma chave de data estrangeira na tabela de fatos para cada marco crítico no processo. Uma linha individual em uma tabela de fatos de snapshot acumulativo, correspondendo, por exemplo, a uma linha em um pedido, é inicialmente inserida quando a linha do pedido é criada. À medida que o progresso do pipeline ocorre, a linha da tabela de fatos de snapshot acumulativo é revisitada e atualizada. Essa atualização consistente das linhas de snapshot acumulativo é única entre os três tipos de tabelas de fatos. Além das chaves de data estrangeiras associadas a cada etapa crítica do processo, as tabelas de fatos de snapshot acumulativo contêm chaves estrangeiras para outras dimensões e, opcionalmente, contêm dimensões degeneradas. Elas frequentemente incluem medidas de lag numéricas consistentes com o grão, juntamente com contadores de conclusão de marcos.

**Tabelas de Fatos Sem Fatos**

Embora a maioria dos eventos de medição capture resultados numéricos, é possível que o evento meramente registre um conjunto de entidades dimensionais se reunindo em um momento no tempo. Por exemplo, um evento de um estudante assistindo a uma aula em um determinado dia pode não ter um fato numérico registrado, mas uma linha de fato com chaves estrangeiras para dia Quado calendário, estudante, professor, localização e aula está bem definida. Da mesma forma, comunicações com clientes são eventos, mas pode não haver métricas associadas. Tabelas de fatos sem fatos também podem ser usadas para analisar o que não aconteceu. Essas consultas sempre têm duas partes: uma tabela de cobertura sem fatos que contém todas as possibilidades de eventos que podem ocorrer e uma tabela de atividade que contém os eventos que realmente ocorreram. Quando a atividade é subtraída da cobertura, o resultado é o conjunto de eventos que não ocorreram.  
  
**Tabelas de Fatos Agregadas ou Cubos OLAP**

Tabelas de fatos agregadas são rollups numéricos simples de dados de tabela de fatos atômicos construídos exclusivamente para acelerar o desempenho das consultas. Essas tabelas de fatos agregadas devem estar disponíveis para a camada de BI ao mesmo tempo que as tabelas de fatos atômicas, para que as ferramentas de BI escolham suavemente o nível agregado apropriado no momento da consulta. Esse processo, conhecido como navegação de agregados, deve ser aberto para que todos os escritores de relatórios, ferramentas de consulta e aplicações de BI colham os mesmos benefícios de desempenho. Um conjunto de agregados bem projetado deve se comportar como índices de banco de dados, que aceleram o desempenho da consulta, mas não são encontrados diretamente pelas aplicações de BI ou usuários de negócios. Tabelas de fatos agregadas contêm chaves estrangeiras para dimensões conformes reduzidas, bem como fatos agregados criados somando medidas de tabelas de fatos mais atômicas. Finalmente, cubos OLAP agregados com medidas resumidas são frequentemente construídos da mesma maneira que agregados relacionais, mas os cubos OLAP são destinados a ser acessados diretamente pelos usuários de negócios.  
  
**Tabelas de Fatos Consolidadas**

Frequentemente é conveniente combinar fatos de múltiplos processos em uma única tabela de fatos consolidada se eles puderem ser expressos no mesmo grão. Por exemplo, vendas reais podem ser consolidadas com previsões de vendas em uma única tabela de fatos para tornar a tarefa de analisar reais versus previsões simples e rápida, em comparação com a montagem de uma aplicação de drill-across usando tabelas de fatos separadas. Tabelas de fatos consolidadas adicionam carga ao processamento ETL, mas facilitam a carga analítica nas aplicações de BI. Elas devem ser consideradas para métricas de processos cruzados que são frequentemente analisadas juntas.

**Técnicas Básicas de Tabelas de Dimensões**

As técnicas nesta seção aplicam-se a todas as tabelas de dimensões. As tabelas de dimensões são discutidas e ilustradas em todos os capítulos.

**Estrutura da Tabela de Dimensões**

Toda tabela de dimensões possui uma única coluna de chave primária. Esta chave primária é incorporada como uma chave estrangeira em qualquer tabela de fatos associada onde o contexto descritivo da linha da dimensão esteja exatamente correto para aquela linha da tabela de fatos. As tabelas de dimensões são geralmente largas, planas e desnormalizadas, com muitos atributos de texto de baixa cardinalidade. Enquanto códigos operacionais e indicadores podem ser tratados como atributos, os atributos de dimensão mais poderosos são preenchidos com descrições detalhadas. Os atributos da tabela de dimensões são o alvo principal de restrições e especificações de agrupamento de consultas e aplicativos de BI. Os rótulos descritivos nos relatórios são tipicamente valores de domínio de atributos de dimensão.

**Chaves Substitutas de Dimensão**

Uma tabela de dimensões é projetada com uma coluna servindo como chave primária única. Esta chave primária não pode ser a chave natural do sistema operacional porque haverá múltiplas linhas de dimensão para aquela chave natural quando as mudanças forem rastreadas ao longo do tempo. Além disso, as chaves naturais para uma dimensão podem ser criadas por mais de um sistema de origem, e essas chaves naturais podem ser incompatíveis ou mal administradas. O sistema DW/BI precisa assumir o controle das chaves primárias de todas as dimensões; em vez de usar chaves naturais explícitas ou chaves naturais com datas anexadas, você deve criar chaves primárias inteiras anônimas para cada dimensão. Essas chaves substitutas de dimensões são números inteiros simples, atribuídos em sequência, começando com o valor 1, sempre que uma nova chave for necessária. A dimensão de data é isenta da regra da chave substituta; essa dimensão altamente previsível e estável pode usar uma chave primária mais significativa.

**Chaves Naturais, Duráveis e Sobrenaturais**

Chaves naturais criadas pelos sistemas de origem operacional estão sujeitas a regras de negócios fora do controle do sistema DW/BI. Por exemplo, um número de empregado (chave natural) pode ser alterado se o empregado se demitir e depois for readmitido. Quando o data warehouse quer ter uma única chave para aquele empregado, uma nova chave durável deve ser criada que seja persistente e não mude nessa situação. Esta chave é às vezes referida como uma chave sobrenatural durável. As melhores chaves duráveis têm um formato independente do processo de negócios original e, portanto, devem ser números inteiros simples atribuídos em sequência a partir de 1. Embora várias chaves substitutas possam ser associadas a um empregado ao longo do tempo à medida que seu perfil muda, a chave durável nunca muda.

**Drill Down**

O drill down é a maneira mais fundamental de analisar dados pelos usuários de negócios. Drill down simplesmente significa adicionar um cabeçalho de linha a uma consulta existente; o novo cabeçalho de linha é um atributo de dimensão anexado à expressão GROUP BY em uma consulta SQL. O atributo pode vir de qualquer dimensão anexada à tabela de fatos na consulta. Drill down não requer a definição de hierarquias ou caminhos de drill down predefinidos.

**Dimensões Degeneradas**

Às vezes, uma dimensão é definida que não tem conteúdo além de sua chave primária. Por exemplo, quando uma fatura tem vários itens de linha, as linhas de fato dos itens de linha herdam todas as chaves estrangeiras das dimensões da fatura, e a fatura fica sem conteúdo único. Mas o número da fatura permanece uma chave de dimensão válida para tabelas de fatos no nível do item de linha. Esta dimensão degenerada é colocada na tabela de fatos com o reconhecimento explícito de que não há tabela de dimensão associada. Dimensões degeneradas são mais comuns com tabelas de fatos de transação e de snapshot acumulativo.

**Dimensões Desnormalizadas e Achatadas**

Em geral, os designers dimensionais devem resistir aos impulsos de normalização causados por anos de designs de bancos de dados operacionais e, em vez disso, desnormalizar as hierarquias de profundidade fixa de muitos para um em atributos separados em uma linha de dimensão achatada. A desnormalização de dimensões apoia os objetivos gêmeos do modelagem dimensional: simplicidade e velocidade.

**Múltiplas Hierarquias em Dimensões**

Muitas dimensões contêm mais de uma hierarquia natural. Por exemplo, dimensões de data de calendário podem ter uma hierarquia de dia para semana para período fiscal, bem como uma hierarquia de dia para mês para ano. Dimensões intensivas em localização podem ter várias hierarquias geográficas. Em todos esses casos, as hierarquias separadas podem coexistir graciosamente na mesma tabela de dimensões.

**Flags e Indicadores como Atributos Textuais**

Abreviações criptográficas, flags de verdadeiro/falso e indicadores operacionais devem ser complementados em tabelas de dimensões com palavras completas que tenham significado quando visualizadas de forma independente. Códigos operacionais com significados embutidos dentro do valor do código devem ser desmembrados, com cada parte do código expandida em seu próprio atributo descritivo separado de dimensão.

**Atributos Nulos em Dimensões**

Atributos de dimensão com valor nulo resultam quando uma determinada linha de dimensão não foi totalmente populada, ou quando há atributos que não são aplicáveis a todas as linhas da dimensão. Em ambos os casos, recomendamos substituir uma string descritiva, como "Desconhecido" ou "Não Aplicável" no lugar do valor nulo. Nulos em atributos de dimensão devem ser evitados porque diferentes bancos de dados lidam com agrupamento e restrição em nulos de maneira inconsistente.

**Dimensões de Data de Calendário**

As dimensões de data de calendário são anexadas a praticamente todas as tabelas de fatos para permitir a navegação da tabela de fatos através de datas familiares, meses, períodos fiscais e dias especiais no calendário. Você nunca gostaria de calcular a Páscoa em SQL, mas sim olhar para ela na dimensão de data de calendário. A dimensão de data de calendário geralmente tem muitos atributos que descrevem características como número da semana, nome do mês, período fiscal e indicador de feriado nacional. Para facilitar a partição, a chave primária de uma dimensão de data pode ser mais significativa, como um número inteiro representando YYYYMMDD, em vez de uma chave substituta atribuída sequencialmente. No entanto, a tabela de dimensão de data precisa de uma linha especial para representar datas desconhecidas ou a serem determinadas. Quando maior precisão é necessária, um carimbo de data/hora separado pode ser adicionado à tabela de fatos. O carimbo de data/hora não é uma chave estrangeira para uma tabela de dimensão, mas sim uma coluna independente. Se os usuários de negócios restringirem ou agruparem por atributos de hora do dia, como agrupamento de parte do dia ou número do turno, você deve adicionar uma chave estrangeira de dimensão de hora do dia separada à tabela de fatos.

**Dimensões de Função**

Uma única dimensão física pode ser referenciada várias vezes em uma tabela de fatos, com cada referência vinculando-se a uma função logicamente distinta para a dimensão. Por exemplo, uma tabela de fatos pode ter várias datas, cada uma das quais representada por uma chave estrangeira para a dimensão de data. É essencial que cada chave estrangeira se refira a uma visualização separada da dimensão de data, de modo que as referências sejam independentes. Essas visualizações separadas da dimensão (com nomes de coluna de atributos exclusivos) são chamadas de funções.

**Dimensões de Lixo**

Processos de negócios transacionais tipicamente produzem uma série de flags e indicadores de baixa cardinalidade. Em vez de fazer dimensões separadas para cada flag e atributo, você pode criar uma única dimensão de lixo combinando-os. Esta dimensão, frequentemente rotulada como uma dimensão de perfil de transação em um esquema, não precisa ser o produto cartesiano de todos os valores possíveis dos atributos, mas deve conter apenas a combinação de valores que realmente ocorrem nos dados de origem.

**Dimensões Flocadas**

Quando um relacionamento hierárquico em uma tabela de dimensões é normalizado, atributos de baixa cardinalidade aparecem como tabelas secundárias conectadas à tabela de dimensão base por uma chave de atributo. Quando este processo é repetido com todas as hierarquias da tabela de dimensões, uma estrutura multinível característica é criada, chamada de floco de neve. Embora o floco de neve represente dados hierárquicos com precisão, você deve evitar flocos de neve porque é difícil para os usuários de negócios entenderem e navegarem pelos flocos de neve. Eles também podem impactar negativamente o desempenho das consultas. Uma tabela de dimensão desnormalizada e achatada contém exatamente as mesmas informações que uma dimensão flocada.

**Dimensões de Suporte**

Uma dimensão pode conter uma referência a outra tabela de dimensões. Por exemplo, uma dimensão de conta bancária pode referenciar uma dimensão separada representando a data de abertura da conta. Essas referências de dimensões secundárias são chamadas de dimensões de suporte. Dimensões de suporte são permitidas, mas devem ser usadas com moderação. Na maioria dos casos, as correlações entre dimensões devem ser rebaixadas para uma tabela de fatos, onde ambas as dimensões são representadas como chaves estrangeiras separadas.

**Integração via Dimensões Conformes**

Um dos sucessos principais da abordagem de modelagem dimensional tem sido definir uma receita simples mas poderosa para integrar dados de diferentes processos de negócios.

**Dimensões Conformes**

As tabelas de dimensões são conformes quando atributos em tabelas de dimensões separadas têm os mesmos nomes de coluna e conteúdos de domínio. Informações de tabelas de fatos separadas podem ser combinadas em um único relatório usando atributos de dimensões conformes que estão associados a cada tabela de fatos. Quando um atributo conforme é usado como cabeçalho de linha (ou seja, a coluna de agrupamento na consulta SQL), os resultados das tabelas de fatos separadas podem ser alinhados nas mesmas linhas em um relatório de drill-across. Esta é a essência da integração em um sistema DW/BI empresarial. Dimensões conformes, definidas uma vez em colaboração com os representantes de governança de dados da empresa, são reutilizadas em tabelas de fatos; elas entregam tanto consistência analítica quanto custos de desenvolvimento reduzidos no futuro, porque a roda não é repetidamente reinventada.

**Dimensões Reduzidas**

Dimensões reduzidas são dimensões conformes que são um subconjunto de linhas e/ou colunas de uma dimensão base. Dimensões reduzidas são necessárias ao construir tabelas de fatos agregadas. Elas também são necessárias para processos de negócios que naturalmente capturam dados em um nível mais alto de granularidade, como uma previsão por mês e marca (em vez da data e produto mais atômicos associados a dados de vendas). Outro caso de subdefinição de dimensões conformes ocorre quando duas dimensões estão no mesmo nível de detalhe, mas uma representa apenas um subconjunto de linhas.

**Drill Across**

Drill across simplesmente significa fazer consultas separadas em duas ou mais tabelas de fatos onde os cabeçalhos de linha de cada consulta consistem em atributos conformes idênticos. Os conjuntos de respostas das duas consultas são alinhados realizando uma operação de sort-merge nos cabeçalhos de linha dos atributos de dimensão comuns. Os fornecedores de ferramentas de BI referem-se a essa funcionalidade por vários nomes, incluindo stitch e consulta multipass.

**Cadeia de Valor**

Uma cadeia de valor identifica o fluxo natural dos principais processos de negócios de uma organização. Por exemplo, a cadeia de valor de um varejista pode consistir em compras para armazenamento para vendas no varejo. A cadeia de valor de um livro-razão geral pode consistir em orçamentação para compromissos para pagamentos. Sistemas de origem operacionais tipicamente produzem transações ou snapshots em cada etapa da cadeia de valor. Como cada processo produz métricas únicas em intervalos de tempo únicos com granularidade e dimensionalidade únicas, cada processo tipicamente gera pelo menos uma tabela de fatos atômica.

**Arquitetura de Barramento de Data Warehouse Empresarial**

A arquitetura de barramento de data warehouse empresarial fornece uma abordagem incremental para construir o sistema DW/BI empresarial. Esta arquitetura divide o processo de planejamento de DW/BI em peças gerenciáveis, focando nos processos de negócios, enquanto entrega integração via dimensões conformes padronizadas que são reutilizadas entre os processos. Ela fornece uma estrutura arquitetônica, enquanto também divide o programa para incentivar implementações ágeis e gerenciáveis correspondentes às linhas na matriz de barramento de data warehouse empresarial. A arquitetura de barramento é independente de tecnologia e plataforma de banco de dados; estruturas dimensionais relacionais e OLAP podem participar.

**Matriz de Barramento de Data Warehouse Empresarial**

A matriz de barramento de data warehouse empresarial é a ferramenta essencial para projetar e comunicar a arquitetura de barramento de data warehouse empresarial. As linhas da matriz são processos de negócios e as colunas são dimensões. As células sombreadas da matriz indicam se uma dimensão está associada a um determinado processo de negócios. A equipe de design verifica cada linha para testar se uma dimensão candidata está bem definida para o processo de negócios e também verifica cada coluna para ver onde uma dimensão deve ser conformada em vários processos de negócios. Além das considerações de design técnico, a matriz de barramento é usada como entrada para priorizar projetos DW/BI com a gestão de negócios, pois as equipes devem implementar uma linha da matriz de cada vez.

**Técnicas de Implementação Detalhada da Matriz de Barramento**

A matriz de barramento de implementação detalhada é uma matriz de barramento mais granular onde cada linha de processo de negócios foi expandida para mostrar tabelas de fatos específicas ou cubos OLAP. Nesse nível de detalhe, a declaração precisa do grão e a lista de fatos podem ser documentadas.

**Matriz de Oportunidades/Stakeholders**

Depois que as linhas da matriz de barramento de data warehouse empresarial foram identificadas, você pode elaborar uma matriz diferente, substituindo as colunas de dimensão por funções de negócios, como marketing, vendas e finanças, e depois sombreando as células da matriz para indicar quais funções de negócios estão interessadas em quais linhas de processos de negócios. A matriz de oportunidades/stakeholders ajuda a identificar quais grupos de negócios devem ser convidados para as sessões de design colaborativo para cada linha centrada em processos.  
  
**Tratando Atributos de Dimensões que Mudam Lentamente**

**Tipo 0: Retenção do Valor Original**

Com o tipo 0, o valor do atributo da dimensão nunca muda, portanto, os fatos são sempre agrupados por este valor original. Tipo 0 é apropriado para qualquer atributo rotulado como "original", como a pontuação de crédito original de um cliente ou um identificador durável. Aplica-se também à maioria dos atributos em uma dimensão de data.

**Tipo 1: Sobrescrever**

Com o tipo 1, o valor antigo do atributo na linha da dimensão é sobrescrito pelo novo valor; os atributos tipo 1 sempre refletem a atribuição mais recente e, portanto, essa técnica destrói o histórico. Embora essa abordagem seja fácil de implementar e não crie linhas de dimensão adicionais, você deve ter cuidado para que as tabelas de fatos agregados e os cubos OLAP afetados por essa mudança sejam recalculados.

**Tipo 2: Adicionar Nova Linha**

Mudanças tipo 2 adicionam uma nova linha na dimensão com os valores de atributos atualizados. Isso requer a generalização da chave primária da dimensão além da chave natural ou durável, pois potencialmente haverá múltiplas linhas descrevendo cada membro. Quando uma nova linha é criada para um membro da dimensão, uma nova chave substituta primária é atribuída e usada como chave estrangeira em todas as tabelas de fatos a partir do momento da atualização até que uma mudança subsequente crie uma nova chave de dimensão e linha de dimensão atualizada. Um mínimo de três colunas adicionais deve ser adicionado à linha da dimensão com mudanças tipo 2: 1) data ou carimbo de data/hora de início da linha; 2) data ou carimbo de data/hora de expiração da linha; e 3) indicador de linha atual.

**Tipo 3: Adicionar Novo Atributo**

Mudanças tipo 3 adicionam um novo atributo na dimensão para preservar o valor antigo do atributo; o novo valor sobrescreve o atributo principal como em uma mudança tipo 1. Esse tipo de mudança tipo 3 é às vezes chamado de realidade alternativa. Um usuário de negócios pode agrupar e filtrar dados de fatos pelo valor atual ou pela realidade alternativa. Essa técnica de dimensão de mudança lenta é usada relativamente com pouca frequência.

**Tipo 4: Adicionar Mini-Dimensão**

A técnica tipo 4 é usada quando um grupo de atributos em uma dimensão muda rapidamente e é separado para uma mini-dimensão. Essa situação é às vezes chamada de dimensão monstro em rápida mudança. Atributos frequentemente usados em tabelas de dimensão com milhões de linhas são candidatos a design de mini-dimensão, mesmo que não mudem frequentemente. A mini-dimensão tipo 4 requer sua própria chave primária única; as chaves primárias da dimensão base e da mini-dimensão são capturadas nas tabelas de fatos associadas.

**Tipo 5: Adicionar Mini-Dimensão e Atributo Tipo 1**

A técnica tipo 5 é usada para preservar com precisão os valores históricos dos atributos, além de relatar fatos históricos de acordo com os valores atuais dos atributos. O tipo 5 se baseia na mini-dimensão tipo 4, incorporando também uma referência atual tipo 1 à mini-dimensão na dimensão base. Isso permite que os atributos da mini-dimensão atualmente atribuídos sejam acessados junto com os outros na dimensão base sem a necessidade de vinculação através de uma tabela de fatos. Logicamente, você representaria a dimensão base e a mini-dimensão de suporte como uma única tabela na área de apresentação. A equipe de ETL deve sobrescrever essa referência tipo 1 da mini-dimensão sempre que a atribuição da mini-dimensão atual mudar.

**Tipo 6: Adicionar Atributos Tipo 1 à Dimensão Tipo 2**

Assim como o tipo 5, o tipo 6 também fornece valores históricos e atuais de atributos de dimensão. O tipo 6 se baseia na técnica tipo 2, incorporando também versões atuais dos mesmos atributos na linha da dimensão, para que as linhas de fatos possam ser filtradas ou agrupadas pelo valor do atributo tipo 2 em vigor quando a medição ocorreu ou pelo valor atual do atributo. Nesse caso, o atributo tipo 1 é sistematicamente sobrescrito em todas as linhas associadas a uma determinada chave durável sempre que o atributo é atualizado.

**Tipo 7: Dimensões Duais Tipo 1 e Tipo 2**

O tipo 7 é a técnica híbrida final usada para suportar relatórios "como era" e "como está". Uma tabela de fatos pode ser acessada através de uma dimensão modelada tanto como uma dimensão tipo 1, mostrando apenas os valores de atributos mais recentes, quanto como uma dimensão tipo 2, mostrando perfis históricos corretos contemporâneos. A mesma tabela de dimensão permite ambas as perspectivas. Tanto a chave durável quanto a chave substituta primária da dimensão são colocadas na tabela de fatos. Para a perspectiva tipo 1, a flag atual na dimensão é restrita para ser atual, e a tabela de fatos é unida via a chave durável. Para a perspectiva tipo 2, a flag atual não é restrita, e a tabela de fatos é unida via a chave substituta primária. Essas duas perspectivas seriam implementadas como visualizações separadas para os aplicativos de BI.  
  
**Tratando Hierarquias de Dimensões**

Hierarquias dimensionais são comuns. Esta seção descreve abordagens para lidar com hierarquias, começando com a mais básica.

**Hierarquias de Profundidade Fixa Posicional**

Uma hierarquia de profundidade fixa é uma série de relações de muitos para um, como produto para marca, para categoria, para departamento. Quando uma hierarquia de profundidade fixa é definida e os níveis de hierarquia têm nomes acordados, os níveis de hierarquia devem aparecer como atributos posicionais separados em uma tabela de dimensão. Uma hierarquia de profundidade fixa é de longe a mais fácil de entender e navegar, desde que os critérios acima sejam atendidos. Também proporciona desempenho de consulta previsível e rápido. Quando a hierarquia não é uma série de relações de muitos para um ou o número de níveis varia de forma que os níveis não têm nomes acordados, uma técnica de hierarquia irregular deve ser usada.

**Hierarquias Ligeiramente Irregulares/Variáveis em Profundidade**

Hierarquias ligeiramente irregulares não têm um número fixo de níveis, mas a faixa de profundidade é pequena. Hierarquias geográficas geralmente variam em profundidade de talvez três níveis a seis níveis. Em vez de usar a complexa maquinaria para hierarquias imprevisivelmente variáveis, você pode forçar hierarquias ligeiramente irregulares em um design posicional de profundidade fixa com atributos de dimensão separados para o número máximo de níveis, e então preencher o valor do atributo com base em regras fornecidas pelos negócios.

**Hierarquias Irregulares/Variáveis em Profundidade com Tabelas de Ponte de Hierarquia**

Hierarquias irregulares de profundidade indeterminada são difíceis de modelar e consultar em um banco de dados relacional. Embora extensões SQL e linguagens de acesso OLAP forneçam algum suporte para relações recursivas de pai/filho, essas abordagens têm limitações. Com extensões SQL, hierarquias irregulares alternativas não podem ser substituídas no momento da consulta, estruturas de propriedade compartilhada não são suportadas e hierarquias irregulares que variam ao longo do tempo não são suportadas. Todas essas objeções podem ser superadas em bancos de dados relacionais modelando uma hierarquia irregular com uma tabela de ponte especialmente construída. Esta tabela de ponte contém uma linha para cada caminho possível na hierarquia irregular e permite que todas as formas de travessia da hierarquia sejam realizadas com SQL padrão, em vez de usar extensões de linguagem especiais.

**Hierarquias Irregulares/Variáveis em Profundidade com Atributos de Caminho**

O uso de uma tabela de ponte para hierarquias irregulares de profundidade variável pode ser evitado implementando um atributo de caminho na dimensão. Para cada linha na dimensão, o atributo de caminho contém uma string de texto especialmente codificada contendo a descrição completa do caminho desde o nó supremo de uma hierarquia até o nó descrito pela linha particular da dimensão. Muitas das solicitações padrão de análise de hierarquia podem então ser tratadas por SQL padrão, sem recorrer a extensões de linguagem SQL. No entanto, a abordagem de caminho não permite substituição rápida de hierarquias alternativas ou hierarquias de propriedade compartilhada. A abordagem de caminho também pode ser vulnerável a mudanças na estrutura da hierarquia irregular que poderiam forçar a rotulação de toda a hierarquia novamente.

**Técnicas Avançadas de Tabelas de Fatos**

As técnicas nesta seção referem-se a padrões menos comuns de tabelas de fatos.

**Chaves Substitutas de Tabelas de Fatos**

Chaves substitutas são usadas para implementar as chaves primárias de quase todas as tabelas de dimensão. Além disso, chaves substitutas de tabela de fatos de coluna única podem ser úteis, embora não sejam necessárias. Chaves substitutas de tabela de fatos, que não estão associadas a nenhuma dimensão, são atribuídas sequencialmente durante o processo de carga de ETL e são usadas 1) como chave primária de coluna única da tabela de fatos; 2) para servir como um identificador imediato de uma linha da tabela de fatos sem navegar por várias dimensões para fins de ETL; 3) para permitir que um processo de carga interrompido seja revertido ou retomado; 4) para permitir que operações de atualização da tabela de fatos sejam decompostas em inserções e exclusões menos arriscadas.

**Tabelas de Fatos Centopeia**

Alguns designers criam dimensões normalizadas separadas para cada nível de uma hierarquia de muitos para um, como uma dimensão de data, dimensão de mês, dimensão de trimestre e dimensão de ano, e então incluem todas essas chaves estrangeiras em uma tabela de fatos. Isso resulta em uma tabela de fatos centopeia com dezenas de dimensões hierarquicamente relacionadas. Tabelas de fatos centopeia devem ser evitadas. Todas essas dimensões hierarquicamente relacionadas de profundidade fixa devem ser colapsadas de volta aos seus grãos únicos mais baixos, como a data no exemplo mencionado. Tabelas de fatos centopeia também resultam quando designers incorporam numerosas chaves estrangeiras para tabelas de dimensão de baixa cardinalidade individuais, em vez de criar uma dimensão de lixo.

**Valores Numéricos como Atributos ou Fatos**

Os designers às vezes encontram valores numéricos que não se encaixam claramente nas categorias de fato ou atributo de dimensão. Um exemplo clássico é o preço de lista padrão de um produto. Se o valor numérico for usado principalmente para fins de cálculo, ele provavelmente pertence à tabela de fatos. Se um valor numérico estável for usado predominantemente para filtragem e agrupamento, ele deve ser tratado como um atributo de dimensão; os valores numéricos discretos podem ser complementados com atributos de faixa de valor (como $0-50). Em alguns casos, é útil modelar o valor numérico tanto como um fato quanto como um atributo de dimensão, como uma métrica quantitativa de entrega no prazo e um descritor textual qualitativo.

**Fatos de Tempo de Espera/Duração**

Tabelas de fatos de snapshot acumulado capturam múltiplos marcos de processo, cada um com uma chave estrangeira de data e possivelmente um carimbo de data/hora. Usuários de negócios frequentemente desejam analisar os tempos de espera ou durações entre esses marcos; às vezes esses tempos de espera são apenas as diferenças entre datas, mas outras vezes os tempos de espera são baseados em regras de negócios mais complicadas. Se houver dezenas de etapas em um pipeline, poderia haver centenas de tempos de espera possíveis. Em vez de forçar a consulta do usuário a calcular cada tempo de espera possível a partir dos carimbos de data/hora ou chaves estrangeiras de dimensão de data, apenas um tempo de espera pode ser armazenado para cada etapa medida em relação ao ponto de início do processo. Então, cada tempo de espera possível entre duas etapas pode ser calculado como uma subtração simples entre os dois tempos de espera armazenados na tabela de fatos.

**Tabelas de Fatos Cabeçalho/Linha**

Sistemas de transação operacionais frequentemente consistem em uma linha de cabeçalho de transação associada a múltiplas linhas de transação. Com esquemas de cabeçalho/linha (também conhecidos como esquemas pai/filho), todas as chaves estrangeiras de nível de cabeçalho e dimensões degeneradas devem ser incluídas na tabela de fatos de nível de linha.

**Fatos Alocados**

É bastante comum em dados de transação de cabeçalho/linha encontrar fatos de granularidade diferente, como uma taxa de frete de cabeçalho. Você deve se esforçar para alocar os fatos do cabeçalho para o nível de linha com base em regras fornecidas pelos negócios, para que os fatos alocados possam ser fatiados e agrupados por todas as dimensões. Em muitos casos, você pode evitar criar uma tabela de fatos de nível de cabeçalho, a menos que essa agregação ofereça vantagens de desempenho de consulta.

**Tabelas de Fatos de Lucros e Perdas Usando Alocações**

Tabelas de fatos que expõem a equação completa de lucro estão entre os entregáveis mais poderosos de um sistema DW/BI empresarial. A equação de lucro é (receita) - (custos) = (lucro). Tabelas de fatos idealmente implementam a equação de lucro no grão da transação de receita atômica e contêm muitos componentes de custo. Como essas tabelas estão no grão atômico, várias agregações são possíveis, incluindo lucratividade do cliente, lucratividade do produto, lucratividade da promoção, lucratividade do canal e outras. No entanto, essas tabelas de fatos são difíceis de construir porque os componentes de custo devem ser alocados de suas fontes originais para o grão da tabela de fatos. Essa etapa de alocação é frequentemente um subsistema ETL importante e uma etapa politicamente carregada que requer apoio executivo de alto nível. Por essas razões, tabelas de fatos de lucros e perdas geralmente não são abordadas nas fases iniciais de implementação de um programa DW/BI.

**Fatos de Múltiplas Moedas**

Tabelas de fatos que registram transações financeiras em múltiplas moedas devem conter um par de colunas para cada fato financeiro na linha. Uma coluna contém o fato expresso na moeda verdadeira da transação, e a outra contém o mesmo fato expresso em uma única moeda padrão que é usada em toda a tabela de fatos. O valor da moeda padrão é criado em um processo ETL de acordo com uma regra de negócios aprovada para conversão de moeda. Esta tabela de fatos também deve ter uma dimensão de moeda para identificar a moeda verdadeira da transação.

**Fatos de Múltiplas Unidades de Medida**

Alguns processos de negócios exigem que os fatos sejam declarados simultaneamente em várias unidades de medida. Por exemplo, dependendo da perspectiva do usuário de negócios, uma cadeia de suprimentos pode precisar relatar os mesmos fatos como paletes, caixas de remessa, caixas de varejo ou unidades de escaneamento individual. Se a tabela de fatos contiver um grande número de fatos, cada um dos quais deve ser expresso em todas as unidades de medida, uma técnica conveniente é armazenar os fatos uma vez na tabela em uma unidade de medida padrão acordada, mas também armazenar simultaneamente fatores de conversão entre a medida padrão e todas as outras. Esta tabela de fatos poderia ser implantada através de visualizações para cada grupo de usuários, usando um fator de conversão selecionado apropriado. Os fatores de conversão devem residir na linha subjacente da tabela de fatos para garantir que o cálculo da visualização seja simples e correto, minimizando a complexidade da consulta.

**Fatos de Ano até a Data**

Usuários de negócios frequentemente solicitam valores de ano até a data (YTD) em uma tabela de fatos. É difícil argumentar contra uma única solicitação, mas solicitações de YTD podem facilmente se transformar em "YTD no fechamento do período fiscal" ou "período fiscal até a data". Uma maneira mais confiável e extensível de lidar com essas solicitações variadas é calcular as métricas de YTD nos aplicativos de BI ou no cubo OLAP, em vez de armazenar fatos de YTD na tabela de fatos.

SQL de Múltiplas Passagens para Evitar Junções de Tabela de Fatos a Tabela de Fatos  
Um aplicativo de BI nunca deve emitir SQL que una duas tabelas de fatos através das chaves estrangeiras da tabela de fatos. É impossível controlar a cardinalidade do conjunto de respostas de tal junção em um banco de dados relacional, e resultados incorretos serão retornados para a ferramenta de BI. Por exemplo, se duas tabelas de fatos contiverem envios e devoluções de produtos de um cliente, essas duas tabelas de fatos não devem ser unidas diretamente através das chaves estrangeiras de cliente e produto. Em vez disso, a técnica de drilling across duas tabelas de fatos deve ser usada, onde os conjuntos de respostas de envios e devoluções são criados separadamente, e os resultados são mesclados por classificação nos valores de atributo de cabeçalho de linha comuns para produzir o resultado correto.

**Acompanhamento de Intervalos de Tempo em Tabelas de Fatos**

Existem três grãos básicos de tabelas de fatos: transação, snapshot periódico e snapshot acumulado. Em casos isolados, é útil adicionar uma data de início de linha, data de expiração de linha e um indicador de linha atual à tabela de fatos, muito parecido com o que você faz com dimensões que mudam lentamente do tipo 2, para capturar um intervalo de tempo quando a linha de fatos estava ativa. Embora seja um padrão incomum, esse padrão aborda cenários como saldos de estoque que mudam lentamente, onde um snapshot periódico frequente carregaria linhas idênticas a cada snapshot.  
  
**Fatos que Chegam Tardiamente**

Uma linha de fatos chega tardiamente se o contexto dimensional mais atual para novas linhas de fatos não corresponder à linha recebida. Isso acontece quando a linha de fatos é atrasada. Nesse caso, as dimensões relevantes devem ser pesquisadas para encontrar as chaves de dimensão que estavam em vigor quando o evento de medição tardio ocorreu.

**Técnicas Avançadas de Dimensão**

As técnicas nesta seção referem-se a padrões mais avançados de tabelas de dimensão.

**Junções de Tabela de Dimensão para Tabela de Dimensão**

As dimensões podem conter referências a outras dimensões. Embora essas relações possam ser modeladas com dimensões de outrigger, em alguns casos, a existência de uma chave estrangeira para a dimensão de outrigger na dimensão base pode resultar em crescimento explosivo da dimensão base, pois mudanças do tipo 2 na dimensão de outrigger forçam o processamento correspondente do tipo 2 na dimensão base. Esse crescimento explosivo pode ser evitado se você rebaixar a correlação entre as dimensões colocando a chave estrangeira do outrigger na tabela de fatos em vez de na dimensão base. Isso significa que a correlação entre as dimensões só pode ser descoberta atravessando a tabela de fatos, mas isso pode ser aceitável, especialmente se a tabela de fatos for um snapshot periódico, onde todas as chaves para todas as dimensões são garantidas para estar presentes para cada período de relatório.

**Dimensões Multivaloradas e Tabelas de Ponte**

Em um esquema dimensional clássico, cada dimensão anexada a uma tabela de fatos tem um único valor consistente com o grão da tabela de fatos. Mas há várias situações em que uma dimensão é legitimamente multivalorada. Por exemplo, um paciente recebendo um tratamento de saúde pode ter múltiplos diagnósticos simultâneos. Nesses casos, a dimensão multivalorada deve ser anexada à tabela de fatos através de uma chave de dimensão de grupo para uma tabela de ponte com uma linha para cada diagnóstico simultâneo em um grupo.

**Tabelas de Ponte Multivaloradas Variáveis no Tempo**

Uma tabela de ponte multivalorada pode precisar ser baseada em uma dimensão de mudança lenta do tipo 2. Por exemplo, a tabela de ponte que implementa a relação muitos-para-muitos entre contas bancárias e clientes individuais geralmente deve ser baseada em dimensões de conta e cliente do tipo 2. Nesse caso, para evitar ligações incorretas entre contas e clientes, a tabela de ponte deve incluir carimbos de data/hora de início e expiração, e o aplicativo solicitante deve restringir a tabela de ponte a um momento específico no tempo para produzir um snapshot consistente.  
  
**Série Temporal de Tags de Comportamento**

Quase todos os textos em um data warehouse são textos descritivos em tabelas de dimensões. Análises de cluster de clientes de data mining normalmente resultam em tags de comportamento textual, muitas vezes identificadas periodicamente. Nesse caso, as medições de comportamento dos clientes ao longo do tempo tornam-se uma sequência dessas tags de comportamento; essa série temporal deve ser armazenada como atributos posicionais na dimensão do cliente, juntamente com uma string de texto opcional para a sequência completa de tags. As tags de comportamento são modeladas em um design posicional porque são o alvo de consultas simultâneas complexas, em vez de cálculos numéricos.

**Grupos de Estudo de Comportamento**

Comportamentos complexos de clientes às vezes só podem ser descobertos por meio de análises iterativas demoradas. Nesses casos, é impraticável incorporar as análises de comportamento em todos os aplicativos de BI que desejam restringir todos os membros da dimensão do cliente que exibem o comportamento complexo. Os resultados das análises de comportamento complexas, no entanto, podem ser capturados em uma tabela simples, chamada grupo de estudo, consistindo apenas das chaves duráveis dos clientes. Esta tabela estática pode então ser usada como uma espécie de filtro em qualquer esquema dimensional com uma dimensão de cliente, restringindo a coluna do grupo de estudo à chave durável da dimensão do cliente no esquema alvo no momento da consulta. Múltiplos grupos de estudo podem ser definidos e grupos de estudo derivados podem ser criados com interseções, uniões e diferenças de conjunto.

**Fatos Agregados como Atributos de Dimensão**

Usuários de negócios frequentemente estão interessados em restringir a dimensão do cliente com base em métricas de desempenho agregadas, como filtrar todos os clientes que gastaram acima de um certo valor em dólar durante o último ano ou talvez ao longo da vida do cliente. Fatos agregados selecionados podem ser colocados em uma dimensão como alvos para restrição e como rótulos de linha para relatórios. As métricas são frequentemente apresentadas como faixas de valor na tabela de dimensão. Atributos de dimensão que representam métricas de desempenho agregadas adicionam carga ao processamento de ETL, mas facilitam a carga analítica na camada de BI.

**Faixas de Valor Dinâmico**

Um relatório de faixas de valor dinâmico é organizado como uma série de cabeçalhos de linha de relatório que definem um conjunto progressivo de intervalos de tamanho variável de um fato numérico alvo. Por exemplo, um relatório comum de faixas de valor em um banco tem muitas linhas com rótulos como "Saldo de 0 a $10", "Saldo de $10,01 a $25", e assim por diante. Este tipo de relatório é dinâmico porque os cabeçalhos de linha específicos são definidos no momento da consulta, não durante o processamento de ETL. As definições de linha podem ser implementadas em uma pequena tabela de dimensão de faixas de valor que é unida através de junções de maior que/menor que à tabela de fatos, ou as definições podem existir apenas em uma instrução SQL CASE. A abordagem da dimensão de faixas de valor provavelmente é de desempenho superior, especialmente em um banco de dados colunar, porque a abordagem de instrução CASE envolve uma varredura quase irrestrita da tabela de fatos.

**Dimensão de Comentários de Texto**

Em vez de tratar comentários livres como métricas textuais em uma tabela de fatos, eles devem ser armazenados fora da tabela de fatos em uma dimensão de comentários separada (ou como atributos em uma dimensão com uma linha por transação se a cardinalidade dos comentários corresponder ao número de transações únicas) com uma chave estrangeira correspondente na tabela de fatos.  
  
**Múltiplos Fusos Horários**

Para capturar tanto o tempo padrão universal quanto os tempos locais em aplicativos de múltiplos fusos horários, chaves estrangeiras duplas devem ser colocadas nas tabelas de fatos afetadas que se unem a duas tabelas de dimensão de data (e potencialmente de horário) com funções específicas.  
  
**Dimensões de Tipo de Medida**

Às vezes, quando uma tabela de fatos tem uma longa lista de fatos que é esparsamente populada em qualquer linha individual, é tentador criar uma dimensão de tipo de medida que colapsa a linha da tabela de fatos para um único fato genérico identificado pela dimensão de tipo de medida. Geralmente não recomendamos essa abordagem. Embora remova todas as colunas de fatos vazias, multiplica o tamanho da tabela de fatos pelo número médio de colunas ocupadas em cada linha e torna os cálculos intra-colunas muito mais difíceis. Esta técnica é aceitável quando o número de fatos potenciais é extremo (nas centenas), mas menos de um punhado seria aplicável a qualquer linha da tabela de fatos.

**Dimensões de Passo**

Processos sequenciais, como eventos de página web, normalmente têm uma linha separada em uma tabela de fatos de transação para cada passo em um processo. Para indicar onde o passo individual se encaixa na sessão geral, é usada uma dimensão de passo que mostra qual número de passo é representado pelo passo atual e quantos mais passos foram necessários para concluir a sessão.  
  
**Dimensões Intercambiáveis a Quente**

Dimensões intercambiáveis a quente são usadas quando a mesma tabela de fatos é alternativamente pareada com diferentes cópias da mesma dimensão. Por exemplo, uma única tabela de fatos contendo cotações de ações pode ser simultaneamente exposta a vários investidores separados, cada um dos quais possui atributos exclusivos e proprietários atribuídos a diferentes ações.  
  
**Dimensões Genéricas Abstratas**

Alguns modeladores são atraídos por dimensões genéricas abstratas. Por exemplo, seus esquemas incluem uma única dimensão de localização genérica em vez de atributos geográficos embutidos nas dimensões de loja, armazém e cliente. Da mesma forma, sua dimensão de pessoa inclui linhas para funcionários, clientes e contatos de fornecedores porque todos são seres humanos, independentemente de atributos significativamente diferentes serem coletados para cada tipo. Dimensões genéricas abstratas devem ser evitadas em modelos dimensionais. Os conjuntos de atributos associados a cada tipo frequentemente diferem. Se os atributos são comuns, como um estado geográfico, então devem ser rotulados de forma única para distinguir o estado da loja do estado do cliente. Finalmente, colocar todas as variedades de locais, pessoas ou produtos em uma única dimensão invariavelmente resulta em uma tabela de dimensão maior. A abstração de dados pode ser apropriada no sistema de origem operacional ou no processamento de ETL, mas impacta negativamente o desempenho e a legibilidade da consulta no modelo dimensional.  
  
**Dimensões de Auditoria**

Quando uma linha da tabela de fatos é criada na retaguarda do ETL, é útil criar uma dimensão de auditoria contendo os metadados de processamento de ETL conhecidos no momento. Uma linha simples da dimensão de auditoria poderia conter um ou mais indicadores básicos de qualidade de dados, talvez derivados do exame de um esquema de evento de erro que registra violações de qualidade de dados encontradas durante o processamento dos dados. Outros atributos úteis de dimensão de auditoria poderiam incluir variáveis de ambiente descrevendo as versões do código ETL usadas para criar as linhas de fatos ou os carimbos de data/hora de execução do processo ETL. Essas variáveis de ambiente são especialmente úteis para fins de conformidade e auditoria porque permitem que as ferramentas de BI investiguem quais linhas foram criadas com quais versões do software ETL.  
  
**Dimensões que Chegam Tardiamente**

Às vezes, os fatos de um processo de negócios operacional chegam minutos, horas, dias ou semanas antes do contexto dimensional associado. Por exemplo, em uma situação de entrega de dados em tempo real, uma linha de redução de inventário pode chegar mostrando a chave natural de um cliente que se comprometeu a comprar um determinado produto. Em um sistema de ETL em tempo real, essa linha deve ser postada na camada de BI, mesmo que a identidade do cliente ou produto não possa ser determinada imediatamente. Nesses casos, linhas de dimensão especiais são criadas com as chaves naturais não resolvidas como atributos. Claro, essas linhas de dimensão devem conter valores genéricos desconhecidos para a maioria das colunas descritivas; presumivelmente, o contexto dimensional adequado virá da fonte em um momento posterior. Quando esse contexto dimensional é finalmente fornecido, as linhas de dimensão de espaço reservado são atualizadas com sobrescritas do tipo 1. Dados de dimensão que chegam tardiamente também ocorrem quando mudanças retroativas são feitas em atributos de dimensão do tipo 2. Nesse caso, uma nova linha precisa ser inserida na tabela de dimensão e, em seguida, as linhas de fatos associadas devem ser restabelecidas.  
  
**Esquemas de Propósito Especial**

Os seguintes padrões de design são necessários para casos de uso específicos.

**Esquemas de Supertipo e Subtipo para Produtos Heterogêneos**

Serviços financeiros e outras empresas frequentemente oferecem uma ampla variedade de produtos em linhas de negócios díspares. Por exemplo, um banco de varejo pode oferecer dezenas de tipos de contas, variando de contas correntes a hipotecas a empréstimos comerciais, mas todas são exemplos de uma conta. Tentativas de construir uma única tabela de fatos consolidada com a união de todos os possíveis fatos, vinculada a tabelas de dimensão com todos os possíveis atributos desses produtos divergentes, falharão porque pode haver centenas de fatos e atributos incompatíveis. A solução é construir uma única tabela de fatos de supertipo que tenha a interseção dos fatos de todos os tipos de contas (junto com uma tabela de dimensão de supertipo contendo os atributos comuns) e, em seguida, sistematicamente construir tabelas de fatos separadas (e tabelas de dimensão associadas) para cada um dos subtipos. Tabelas de fatos de supertipo e subtipo também são chamadas de tabelas de fatos principais e personalizadas.  
  
**Tabelas de Fatos em Tempo Real**

Tabelas de fatos em tempo real precisam ser atualizadas com mais frequência do que o processo de lote noturno mais tradicional. Existem muitas técnicas para suportar essa necessidade, dependendo das capacidades do DBMS ou do cubo OLAP usado para a implantação final na camada de relatórios de BI. Por exemplo, uma "partição quente" pode ser definida em uma tabela de fatos que é fixada na memória física. Agregações e índices são deliberadamente não construídos nessa partição. Outros DBMSs ou cubos OLAP podem suportar atualizações adiadas que permitem que consultas existentes sejam executadas até a conclusão, mas depois realizem as atualizações.  
  
**Esquemas de Eventos de Erro**

Gerenciar a qualidade dos dados em um data warehouse requer um sistema abrangente de filtros ou telas de qualidade de dados que testem os dados à medida que fluem dos sistemas de origem para a plataforma de BI. Quando uma tela de qualidade de dados detecta um erro, este evento é registrado em um esquema dimensional especial que está disponível apenas na retaguarda do ETL. Este esquema consiste em uma tabela de fatos de eventos de erro cujo grão é o evento de erro individual e uma tabela de fatos de detalhes de eventos de erro associada cujo grão é cada coluna em cada tabela que participa de um evento de erro.

**Capítulo 17**

**Visão Geral do Ciclo de Vida do Kimball DW/BI**

A mudança de foco é bastante dramática neste capítulo. Em vez de nos concentrarmos nas técnicas de modelagem dimensional de Kimball, voltamos nossa atenção para tudo o que ocorre durante o curso de um projeto de design e implementação de data warehouse/inteligência de negócios (DW/BI). Neste capítulo, cobriremos a vida de um projeto DW/BI desde a sua criação até a manutenção contínua, identificando as melhores práticas em cada etapa, bem como as vulnerabilidades potenciais. Uma cobertura mais abrangente do Ciclo de Vida de Kimball está disponível em The Data Warehouse Lifecycle Toolkit, Second Edition por Ralph Kimball, Margy Ross, Warren Thornthwaite, Joy Mundy e Bob Becker (Wiley, 2008). Este capítulo é um curso intensivo baseado no texto completo, que pesa mais de 600 páginas.

Você pode perceber que o conteúdo deste capítulo é aplicável apenas aos gerentes de projetos DW/BI, mas pensamos de forma diferente. Implementar um sistema DW/BI requer atividades integradas de maneira rigorosa. Acreditamos que todos na equipe do projeto, incluindo analistas, arquitetos, designers e desenvolvedores, precisam de uma compreensão de alto nível do ciclo de vida completo.

Este capítulo fornece uma visão geral da abordagem completa do Ciclo de Vida Kimball; recomendações específicas sobre modelagem dimensional e tarefas de ETL são adiadas para os capítulos subsequentes. Vamos nos aprofundar no processo de workshop de modelagem colaborativa no Capítulo 18: Dimensional Modeling Process and Tasks, e depois fazer uma imersão similar nas atividades de ETL no Capítulo 20: ETL System Design and Development Process and Tasks.  
  
O Capítulo 17 aborda os seguintes conceitos:

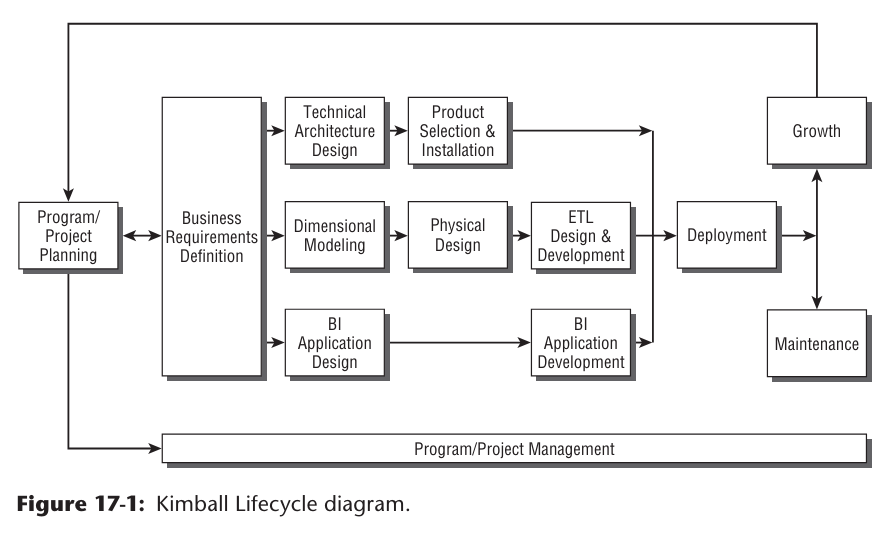
* Orientação sobre o Ciclo de Vida Kimball
* Planejamento e gerenciamento de programa/projeto DW/BI
* Táticas para coleta de requisitos de negócios, incluindo priorização
* Processo para desenvolver a arquitetura técnica e selecionar produtos
* Considerações sobre design físico, incluindo agregação e indexação
* Atividades de design e desenvolvimento de aplicativos BI

**Recomendações para implantação, manutenção contínua e crescimento futuro**

**Roteiro do Ciclo de Vida**

Quando dirigimos para um lugar onde nunca estivemos antes, a maioria de nós confia em um roteiro, embora exibido via GPS. Da mesma forma, um roteiro é extremamente útil se estivermos prestes a embarcar na jornada desconhecida de data warehousing e business intelligence. Os autores de The Data Warehouse Lifecycle Toolkit basearam-se em décadas de experiência para desenvolver a abordagem do Ciclo de Vida Kimball. Quando introduzimos o Ciclo de Vida pela primeira vez em 1998, chamamos de Business Dimensional Lifecycle, um nome que reforçava nossos princípios-chave para o sucesso em data warehouse: focar nas necessidades do negócio, apresentar dados estruturados dimensionalmente aos usuários e enfrentar projetos iterativos e gerenciáveis. Na década de 1990, éramos uma das poucas organizações que enfatizavam esses princípios fundamentais, então o nome diferenciava nossos métodos dos outros. Ainda estamos firmemente ligados a esses princípios, que desde então se tornaram práticas recomendadas geralmente aceitas na indústria, mas renomeamos nossa abordagem para Kimball Lifecycle porque é assim que a maioria das pessoas se refere a ela.

A abordagem geral do Ciclo de Vida Kimball é encapsulada na Figura 17-1. O diagrama ilustra a sequência de tarefas, dependência e simultaneidade. Serve como um roteiro para ajudar as equipes a fazer a coisa certa na hora certa. O diagrama não reflete uma linha do tempo absoluta; embora as caixas tenham larguras iguais, há uma vasta diferença no tempo e esforço necessários para cada atividade principal.

  
**NOTA:** Dada a recente ênfase da indústria em metodologias ágeis, queremos lembrar aos leitores sobre a discussão do tópico no Capítulo 1: Data Warehousing, Business Intelligence, and Dimensional Modeling Primer. A abordagem do Ciclo de Vida Kimball e as metodologias ágeis compartilham algumas doutrinas comuns: foco no valor do negócio, colaboração com o negócio e desenvolvimento incremental. No entanto, também acreditamos fortemente que o design e desenvolvimento do sistema DW/BI precisam ser construídos sobre uma arquitetura de dados sólida e uma base de governança, orientada pela arquitetura de barramento. Também acreditamos que a maioria das situações justifica o agrupamento de múltiplas entregas ágeis em uma versão mais completa antes de serem amplamente implantadas na comunidade empresarial.

**Marco do Roteiro**

Antes de mergulhar em detalhes, reserve um momento para se orientar no roteiro. O Ciclo de Vida começa com o planejamento do programa/projeto, como seria de esperar. Este módulo avalia a prontidão da organização para uma iniciativa DW/BI, estabelece o escopo preliminar e a justificativa, obtém recursos e lança o programa/projeto. O gerenciamento contínuo do projeto serve como base para manter as atividades restantes nos trilhos.

A segunda tarefa principal na Figura 17-1 concentra-se na definição de requisitos de negócios. Há uma seta bidirecional entre o planejamento do programa/projeto e a definição de requisitos de negócios devido à interação entre essas atividades. Alinhar a iniciativa DW/BI com os requisitos de negócios é absolutamente crucial. Tecnologias de ponta não salvarão um ambiente DW/BI que não foca no negócio. Os usuários de negócios e seus requisitos têm impacto em quase todas as decisões de design e implementação feitas durante o curso de um projeto DW/BI. No roteiro da Figura 17-1, isso é refletido pelas três trilhas paralelas que seguem.

A trilha superior da Figura 17-1 lida com tecnologia. O design da arquitetura técnica estabelece a estrutura geral para suportar a integração de várias tecnologias. Usando as capacidades identificadas no design da arquitetura como uma lista de compras, você então avalia e seleciona produtos específicos. Note que a seleção de produtos não é a primeira caixa no roteiro. Um dos erros mais frequentes cometidos por equipes novatas é selecionar produtos sem uma compreensão clara do que estão tentando realizar. Isso é semelhante a pegar um martelo, independentemente de você precisar pregar um prego ou apertar um parafuso.

A trilha do meio emanando da definição de requisitos de negócios concentra-se em dados. Começa traduzindo os requisitos em um modelo dimensional, como temos praticado. O modelo dimensional é então transformado em uma estrutura física. O foco está em estratégias de otimização de desempenho, como agregação, indexação e particionamento, durante o design físico. Por último, mas não menos importante, o sistema de ETL é projetado e desenvolvido. Como mencionado anteriormente, as caixas de tamanho igual não representam esforços de tamanho igual; isso se torna óbvio com a diferença de carga de trabalho entre o design físico e as atividades exigentes centradas no ETL.

O conjunto final de tarefas originadas dos requisitos de negócios é o design e desenvolvimento dos aplicativos de BI. O projeto DW/BI não está concluído quando você entrega dados. Aplicativos de BI, na forma de templates parametrizados e análises, satisfarão uma grande porcentagem das necessidades analíticas dos usuários de negócios.

As trilhas de tecnologia, dados e aplicativos de BI, junto com uma dose saudável de educação e suporte, convergem para uma implantação bem orquestrada. A partir daí, é necessária manutenção contínua para garantir que o sistema DW/BI permaneça saudável. Finalmente, você lida com o crescimento futuro iniciando projetos subsequentes, cada um retornando ao início do Ciclo de Vida novamente.

Agora que você tem uma compreensão de alto nível do roteiro geral, descreveremos cada uma das caixas na Figura 17-1 em mais detalhes.

**Atividades de Lançamento do Ciclo de Vida**

As seções a seguir descrevem as melhores práticas e armadilhas a serem evitadas ao iniciar um projeto DW/BI.

**Planejamento e Gerenciamento do Programa/Projeto**

Não surpreendentemente, a iniciativa DW/BI começa com uma série de atividades de planejamento de programa e projeto.

**Avaliando a Prontidão**

Antes de prosseguir com um esforço DW/BI, é prudente tirar um momento para avaliar a prontidão da organização para avançar. Com base em nossa experiência acumulada de centenas de compromissos com clientes, três fatores diferenciam projetos que foram predominantemente tranquilos daqueles que implicaram uma luta constante. Esses fatores são indicadores principais de sucesso em DW/BI; descreveremos as características em ordem de importância.

O fator de prontidão mais crítico é ter um forte patrocinador executivo de negócios. Patrocinadores de negócios devem ter uma visão clara do potencial impacto do sistema DW/BI na organização. Idealmente, patrocinadores de negócios têm um histórico de sucesso com outras iniciativas internas. Eles devem ser líderes politicamente astutos que possam convencer seus pares a apoiar o esforço. É um cenário muito mais arriscado se o diretor de informações (CIO) for o patrocinador designado; preferimos muito mais o compromisso visível de um parceiro de negócios.

O segundo fator de prontidão é ter uma motivação comercial forte e convincente para enfrentar a iniciativa DW/BI. Este fator muitas vezes anda de mãos dadas com o patrocínio. O projeto DW/BI precisa resolver problemas críticos de negócios para obter os recursos necessários para um lançamento bem-sucedido e uma vida útil saudável. Motivação convincente geralmente cria um senso de urgência, seja a motivação de fontes externas, como fatores competitivos, ou fontes internas, como a incapacidade de analisar o desempenho de toda a organização após aquisições.

O terceiro fator ao avaliar a prontidão é a viabilidade. Existem vários aspectos da viabilidade, incluindo viabilidade técnica e de recursos, mas a viabilidade dos dados é a mais crucial. Você está coletando dados reais em sistemas de origem operacional reais para apoiar os requisitos de negócios? A viabilidade dos dados é uma grande preocupação porque não há uma solução de curto prazo se você não estiver coletando dados de origem razoavelmente limpos no nível de granularidade certo.

**Definindo o Escopo e Justificativa**

Quando você estiver confortável com a prontidão da organização, é hora de estabelecer limites em torno de um projeto inicial. A definição do escopo requer a contribuição conjunta da organização de TI e da gestão de negócios. O escopo de um projeto DW/BI deve ser tanto significativo para a organização de negócios quanto gerenciável para a organização de TI.

Inicialmente, você deve abordar projetos que se concentram em dados de um único processo de negócios; guarde os projetos mais desafiadores, de processos cruzados, para uma fase posterior. Lembre-se de evitar a Lei do Demais ao definir o escopo—um prazo muito curto para um projeto com muitos sistemas de origem e muitos usuários em muitos locais com requisitos analíticos muito diversos.

A justificativa requer uma estimativa dos benefícios e custos associados à iniciativa DW/BI. Esperançosamente, os benefícios esperados superam amplamente os custos. O TI geralmente é responsável por calcular as despesas. Os sistemas DW/BI tendem a se expandir rapidamente, então, certifique-se de que as estimativas permitam espaço para crescimento de curto prazo. Ao contrário do desenvolvimento de sistemas operacionais, onde os requisitos de recursos diminuem após a produção, as necessidades de suporte contínuo de DW/BI não diminuirão significativamente ao longo do tempo.

A comunidade de negócios deve ter a responsabilidade principal de determinar os benefícios financeiros esperados. Os ambientes DW/BI geralmente são justificados com base em oportunidades de aumento de receita ou lucro, em vez de focar apenas na redução de despesas. Fornecer "uma única versão da verdade" ou "acesso flexível à informação" não é justificativa financeira suficiente. Você precisa aprofundar para determinar o impacto quantificável da melhoria na tomada de decisões possibilitada por essas frases de efeito. Se você estiver lutando com a justificativa, provavelmente isso é um sintoma de que a iniciativa está focada no patrocinador de negócios ou no problema errado.

**Montando a Equipe**

Os projetos DW/BI exigem a integração de uma equipe multifuncional com recursos tanto da comunidade de negócios quanto de TI. É comum que a mesma pessoa ocupe vários papéis na equipe; a designação de recursos nomeados para funções depende da magnitude e do escopo do projeto, bem como da disponibilidade, capacidade e experiência do indivíduo.

**Do lado dos negócios, precisaremos de representantes para preencher os seguintes papéis:**

**Patrocinador de negócios:** O patrocinador é o cliente final do sistema DW/BI, bem como seu maior defensor. O patrocínio às vezes assume a forma de um comitê executivo, especialmente para iniciativas de toda a empresa.

**Responsável de negócios:** Em uma organização grande, o patrocinador pode estar muito distante ou inacessível para a equipe do projeto. Nesse caso, o patrocinador às vezes delega responsabilidades menos estratégicas do DW/BI a um gerente intermediário na organização. Este responsável deve possuir as mesmas características do patrocinador.

**Líder de negócios:** O líder do projeto de negócios é uma pessoa respeitada que está altamente envolvida no projeto, comunicando-se com o gerente de projeto diariamente. Às vezes, o responsável de negócios assume esse papel.

**Usuários de negócios:** Idealmente, os usuários de negócios são os fãs entusiásticos do ambiente DW/BI. Você precisa envolvê-los cedo e frequentemente, começando com o escopo do projeto e requisitos de negócios. A partir daí, você deve encontrar maneiras criativas de manter o interesse e a participação deles ao longo do projeto. Lembre-se, o envolvimento dos usuários de negócios é crítico para a aceitação do DW/BI. Sem usuários de negócios, o sistema DW/BI é um exercício técnico em vão.

Várias posições são preenchidas tanto pelas organizações de negócios quanto de TI. Esses mediadores podem ser recursos técnicos que entendem o negócio ou recursos de negócios que entendem tecnologia:

**Analista de negócios:** Essa pessoa é responsável por determinar as necessidades de negócios e traduzi-las em requisitos arquitetônicos, de dados e de aplicativos de BI.

**Administrador de dados:** Este especialista em assunto é frequentemente o recurso atual para análises ad hoc. Eles entendem o que os dados significam, como são usados e onde as inconsistências de dados estão escondidas. Dada a necessidade de consenso organizacional em torno dos dados dimensionais principais, este pode ser um papel politicamente desafiador, como descrevemos no Capítulo 4: Inventário.

**Designer/desenvolvedor de aplicativos de BI:** Os recursos de aplicativos de BI são responsáveis por projetar e desenvolver o conjunto inicial de templates analíticos, bem como fornecer suporte contínuo aos aplicativos de BI.

**Os seguintes papéis são geralmente preenchidos pela organização de TI:**

**Gerente de projeto:** O gerente de projeto é uma posição crítica. Esta pessoa deve estar confortável e ser respeitada pelos executivos de negócios, bem como pelos recursos técnicos. As habilidades de comunicação e gerenciamento de projetos do gerente de projeto devem ser excelentes.

**Arquiteto técnico:** O arquiteto é responsável pela arquitetura técnica geral. Esta pessoa desenvolve o plano que integra as funcionalidades técnicas necessárias e ajuda a avaliar produtos com base na arquitetura geral.

**Arquiteto/modelador de dados:** Este recurso provavelmente vem de um background de dados transacionais com forte ênfase na normalização. Esta pessoa deve abraçar os conceitos de modelagem dimensional e ser empática aos requisitos do negócio, em vez de focar estritamente em economizar espaço ou reduzir a carga de trabalho do ETL.

**Administrador de banco de dados:** Assim como o modelador de dados, o administrador de banco de dados deve estar disposto a abandonar alguns dos truísmos tradicionais de administração de banco de dados, como ter apenas um índice em uma tabela relacional.

**Coordenador de metadados:** Esta pessoa ajuda a estabelecer a estratégia de repositório de metadados e garante que os metadados apropriados sejam coletados, gerenciados e disseminados.

**Arquiteto/designer de ETL:** Este papel é responsável pelo design do ambiente e dos processos de ETL.

**Desenvolvedor de ETL:** Com base na direção do arquiteto/designer de ETL, o desenvolvedor constrói e automatiza os processos, provavelmente usando uma ferramenta de ETL.

Queremos salientar novamente que esta é uma lista de papéis, não de pessoas. Especialmente em lojas menores, indivíduos talentosos preencherão muitos desses papéis simultaneamente.

**Desenvolvendo e Mantendo o Plano**

O plano do projeto DW/BI identifica todas as tarefas necessárias do Ciclo de Vida. Uma lista detalhada de tarefas está disponível no site do Kimball Group em www.kimballgroup.com; confira a guia Tools & Utilities sob o título do livro The Data Warehouse Lifecycle Toolkit, Second Edition.

Qualquer bom gerente de projeto sabe que os principais membros da equipe devem desenvolver estimativas do esforço necessário para suas tarefas; o gerente de projeto não pode ditar o tempo permitido e esperar conformidade. O plano do projeto deve identificar pontos de verificação de aceitação com representantes de negócios após cada marco e entrega principal do roteiro para garantir que o projeto permaneça no caminho certo.

Os projetos DW/BI exigem comunicação ampla. Embora os gerentes de projeto geralmente se sobressaiam nas comunicações intra-equipe, eles também devem estabelecer uma estratégia de comunicação descrevendo a frequência, o fórum e as principais mensagens para outras partes interessadas, incluindo os patrocinadores de negócios, a comunidade de negócios e outros colegas de TI.

Finalmente, os projetos DW/BI são vulneráveis ao escopo ampliado, principalmente devido à forte necessidade de satisfazer os requisitos dos usuários de negócios. Você tem várias opções quando confrontado com mudanças: aumentar o escopo (adicionando tempo, recursos ou orçamento), jogar o jogo de soma zero (mantendo o escopo original, desistindo de algo em troca) ou dizer "não" (sem realmente dizer "não", tratando a mudança como uma solicitação de melhoria). O mais importante sobre decisões de escopo é que elas não devem ser tomadas em um vácuo de TI. A resposta certa depende da situação. Agora é a hora de aproveitar a parceria com o negócio para chegar a uma resposta que todos possam viver.

**Definição de Requisitos de Negócios**

Colaborar com os usuários de negócios para entender seus requisitos e garantir seu comprometimento é absolutamente essencial para o sucesso de data warehousing e business intelligence. Esta seção enfoca técnicas básicas para coleta de requisitos de negócios.

**Planejamento Prévio de Requisitos**

Antes de se sentar com representantes de negócios para coletar seus requisitos, sugerimos o seguinte para garantir sessões produtivas:

**Escolha do Fórum**

As sessões de requisitos de usuários de negócios são tipicamente intercaladas com sessões de descoberta de dados de especialistas em sistemas de origem. Esta abordagem dupla fornece insights sobre as necessidades do negócio com as realidades dos dados. No entanto, você não pergunta aos representantes de negócios sobre a granularidade ou dimensionalidade de seus dados críticos. Você precisa falar com eles sobre o que fazem, por que fazem, como tomam decisões e como esperam tomar decisões no futuro. Como terapia organizacional, você está tentando detectar os problemas e oportunidades.

**Existem duas técnicas principais para coleta de requisitos:** entrevistas ou sessões facilitadas. Ambas têm suas vantagens e desvantagens. As entrevistas incentivam a participação individual e também são mais fáceis de agendar. Sessões facilitadas podem reduzir o tempo decorrido para coletar requisitos, mas exigem mais compromisso de tempo de cada participante.

Com base em nossa experiência, pesquisas não são uma ferramenta razoável para coleta de requisitos porque são planas e bidimensionais. Os respondentes auto-selecionados respondem apenas às perguntas que pensamos em fazer antecipadamente; não há opção de aprofundar mais. Além disso, os instrumentos de pesquisa não ajudam a forjar o vínculo entre os usuários de negócios e a iniciativa DW/BI que buscamos.

Geralmente usamos uma abordagem híbrida com entrevistas para coletar os detalhes e, em seguida, facilitação para levar o grupo ao consenso. Embora descrevamos essa abordagem híbrida em mais detalhes, grande parte da discussão se aplica também à facilitação pura.

A escolha do fórum de coleta de requisitos depende das habilidades da equipe, da cultura da organização e do que os usuários de negócios já foram submetidos. Um tamanho definitivamente não serve para todos.

**Identificação e Preparação da Equipe de Requisitos**

Independentemente da abordagem, você precisa identificar e preparar os membros da equipe do projeto envolvidos. Se você estiver fazendo entrevistas, precisa identificar um entrevistador principal cuja responsabilidade principal é fazer ótimas perguntas abertas. Enquanto isso, o escriba da entrevista toma notas copiosas. Embora um gravador possa fornecer uma cobertura mais completa de cada entrevista, não usamos um porque muda a dinâmica da reunião. Nossa preferência é ter uma segunda pessoa na sala com outro cérebro e um conjunto de olhos e ouvidos, em vez de depender da tecnologia. Frequentemente convidamos um ou dois membros adicionais do projeto (dependendo do número de entrevistados) como observadores, para que possam ouvir diretamente as opiniões dos usuários.

Antes de se sentar com os usuários de negócios, você precisa garantir que está abordando as sessões com a mentalidade correta. Não presuma que já sabe tudo; você definitivamente aprenderá mais sobre o negócio durante as sessões. Por outro lado, você deve fazer algum dever de casa pesquisando fontes disponíveis, como o relatório anual, site e organograma interno.

Como a chave para obter as respostas certas é fazer as perguntas certas, recomendamos elaborar questionários. O questionário não deve ser visto como um roteiro; é uma ferramenta para organizar seus pensamentos e servir como um dispositivo de fallback caso sua mente fique em branco durante a sessão. O questionário será atualizado ao longo do processo de entrevista à medida que a equipe se familiariza mais com o assunto do negócio.

**Seleção, Agendamento e Preparação dos Representantes de Negócios**

Se esta for sua primeira incursão em DW/BI, ou um esforço para desenvolver uma estratégia coesa para lidar com silos de dados existentes, você deve conversar com pessoas de negócios que representam um espectro horizontal razoável da organização. Esta cobertura é crítica para formular o blueprint da matriz de barramento do data warehouse empresarial. Você precisa entender os dados e vocabulário comuns em funções de negócios essenciais para construir um ambiente extensível.

Dentro da comunidade de usuários-alvo, você deve cobrir verticalmente a organização. As equipes de projetos DW/BI naturalmente gravitam em direção aos analistas de negócios poderosos. Embora sua visão seja valiosa, você não pode ignorar executivos seniores e a gerência intermediária. Caso contrário, você corre o risco de se concentrar excessivamente no tático aqui e agora e perder de vista a direção estratégica do grupo.

Agendar os representantes de negócios pode ser a tarefa de requisitos mais onerosa; seja especialmente gentil com os assistentes administrativos do departamento. Preferimos nos encontrar com executivos individualmente. Reuniões com um grupo homogêneo de duas a três pessoas são apropriadas para aqueles mais abaixo na hierarquia organizacional. Permita 1 hora para reuniões individuais e 1 hora e meia para pequenos grupos. O agendador precisa permitir meia hora entre reuniões para o debriefing e outras necessidades. A entrevista é extremamente cansativa devido ao foco necessário. Consequentemente, não agende mais de três a quatro sessões em um dia.

Quando se trata de preparar os entrevistados, o patrocinador de negócios deve comunicar-se com eles, enfatizando seu comprometimento com o esforço e a importância da participação de todos. Os entrevistados devem ser solicitados a trazer cópias de seus relatórios e análises principais para a sessão. Esta comunicação dissemina uma mensagem consistente sobre o projeto, além de transmitir a propriedade do negócio sobre a iniciativa. Ocasionalmente, os entrevistados relutam em trazer os "coroas" de dados do negócio para a reunião, especialmente com um consultor externo. No entanto, quase sempre descobrimos que essas pessoas correrão entusiasticamente de volta aos seus escritórios no final da entrevista para trazer de volta esses mesmos relatórios.

**Coleta de Requisitos de Negócios**

É hora de se sentar cara a cara para coletar os requisitos de negócios. O processo geralmente flui de uma introdução por meio de questionamento estruturado até uma conclusão final.

**Lançamento**

A responsabilidade por introduzir a sessão deve ser estabelecida antes de se reunir na sala de conferências. A pessoa designada para o pontapé inicial deve roteirizar os principais pontos de discussão para os primeiros minutos, quando o tom da reunião de entrevista é estabelecido. A introdução deve transmitir uma mensagem clara e centrada nos negócios e não divagar com hardware, software e outros jargões técnicos.

**Fluxo da Entrevista**

O objetivo de uma entrevista é fazer os usuários de negócios falarem sobre o que fazem e por que fazem. Um lugar simples e não ameaçador para começar é perguntar sobre responsabilidades de trabalho e ajuste organizacional. Isso é uma bola fácil que os entrevistados podem responder facilmente. A partir daí, você normalmente pergunta sobre suas principais métricas de desempenho. Determinar como eles acompanham o progresso e o sucesso se traduz diretamente no modelo dimensional; eles estão lhe contando sobre seus principais processos de negócios e fatos sem que você faça essas perguntas diretamente.

Se você se encontrar com uma pessoa que tem mais experiência prática com dados, deve sondar para entender melhor a dimensionalidade do negócio. Perguntas como "Como você distingue entre produtos (ou agentes, provedores ou instalações)?" ou "Como você categoriza naturalmente os produtos?" ajudam a identificar atributos dimensionais e hierarquias chave. Se o entrevistado for mais analítico, pergunte sobre os tipos de análise atualmente gerados.

Entender a natureza dessas análises e se elas são ad hoc ou padronizadas fornece informações sobre os requisitos da ferramenta de BI, bem como sobre o processo de design de aplicativos de BI. Esperançosamente, o entrevistado trouxe cópias de planilhas e relatórios chave. Em vez de guardá-los em uma pasta, é útil entender como o entrevistado usa a análise hoje, bem como oportunidades de melhoria. Ao contrário do conselho de alguns especialistas da indústria, você não pode projetar um ambiente analítico extensível apenas fazendo com que os usuários concordem com seus cinco principais relatórios. As perguntas dos usuários certamente mudarão; consequentemente, você deve resistir à tentação de restringir seu foco de design a um suposto top cinco.

Se você se encontrar com executivos de negócios, não mergulhe nesses detalhes táticos. Em vez disso, pergunte-lhes sobre sua visão para aproveitar melhor a informação em toda a organização. Talvez a equipe do projeto esteja imaginando um ambiente totalmente ad hoc, enquanto a gestão de negócios está mais interessada na entrega de análises padronizadas. Você precisa garantir que a entrega do DW/BI corresponda à demanda e expectativas do negócio.

Pergunte a cada entrevistado sobre o impacto do acesso aprimorado à informação. Você provavelmente já recebeu financiamento preliminar para o projeto, mas nunca é demais capturar mais benefícios potenciais quantificáveis.

**Conclusão**

À medida que a entrevista chega ao fim, pergunte a cada entrevistado sobre seus critérios de sucesso para o projeto. Claro, cada critério deve ser mensurável. "Fácil de usar" e "rápido" significam algo diferente para cada pessoa, então os entrevistados precisam articular especificidades, como suas expectativas em relação à quantidade de treinamento necessária para executar relatórios de BI predefinidos.

Neste ponto, sempre faça um aviso amplo. Os entrevistados devem entender que, só porque você discutiu uma capacidade na reunião, isso não garante que será incluída na primeira fase do projeto. Agradeça aos entrevistados por suas ideias brilhantes e informe-os sobre o que acontecerá a seguir e qual será o envolvimento deles.

**Conduzindo Entrevistas Centradas em Dados**

Enquanto nos concentramos em entender os requisitos de negócios, é útil intercalar sessões com os gurus de dados dos sistemas de origem ou especialistas em assuntos para avaliar a viabilidade de apoiar as necessidades de negócios. Essas entrevistas centradas em dados são bastante diferentes das descritas anteriormente. O objetivo é verificar se os dados principais necessários existem antes que o impulso cresça por trás dos requisitos. Nessas entrevistas centradas em dados, você pode ir tão longe a ponto de pedir alguns resultados iniciais de perfil de dados, como valores de domínio e contagens de alguns campos de dados críticos, para serem fornecidos posteriormente, apenas para garantir que você não está em terreno instável. Uma auditoria de dados mais completa ocorrerá durante o processo de modelagem dimensional. Tente aprender o suficiente neste ponto para gerenciar adequadamente as expectativas da organização.

**Documentando Requisitos**

Imediatamente após a entrevista, a equipe da entrevista deve fazer um debriefing. Você deve garantir que todos estejam na mesma página sobre o que foi aprendido. Também é útil se todos revisarem suas anotações logo após a sessão para preencher lacunas enquanto a entrevista ainda está fresca. Abreviações e frases parciais nas anotações tornam-se incompreensíveis após alguns dias! Da mesma forma, examine os relatórios coletados para obter mais informações sobre a dimensionalidade que deve ser suportada no data warehouse.

Neste ponto, é hora de documentar o que você ouviu. Embora a documentação seja a atividade menos favorita de todos, é crítica tanto para validação do usuário quanto para materiais de referência da equipe do projeto. Existem dois níveis potenciais de documentação resultantes do processo de requisitos. O primeiro é escrever cada entrevista individual; esta atividade é demorada porque a redação não deve ser apenas uma transcrição do fluxo de consciência, mas deve fazer sentido para alguém que não estava na entrevista. A documentação mais crítica é um documento consolidado de conclusões. Este documento é organizado em torno dos principais processos de negócios. Como você aborda projetos DW/BI em uma base de processo por processo, é apropriado estruturar os requisitos de negócios nos mesmos compartimentos que, por sua vez, se tornarão esforços de implementação.

Ao escrever o documento de conclusões, você deve começar com um resumo executivo, seguido de uma visão geral do projeto cobrindo o processo utilizado e os participantes envolvidos. A maior parte do documento centra-se nos processos de negócios; para cada processo, descreva por que os usuários de negócios querem analisar as métricas de desempenho do processo, quais capacidades eles querem, suas limitações atuais e os benefícios ou impactos potenciais. Comentários sobre a viabilidade de enfrentar cada processo também são importantes.

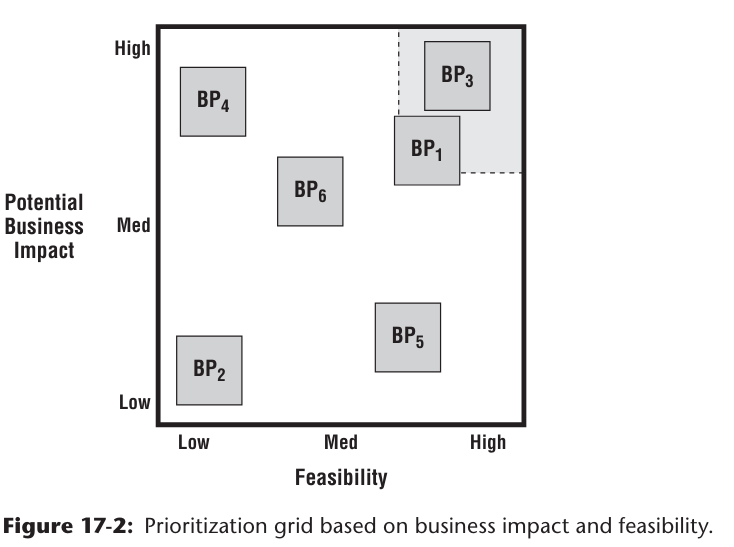
Conforme descrito no Capítulo 4 e ilustrado na Figura 4-11, os processos às vezes são revelados em uma matriz de oportunidade/partes interessadas para transmitir o impacto em toda a organização. Nesse caso, as linhas da matriz de oportunidades identificam processos de negócios, assim como uma matriz de barramento. No entanto, na matriz de oportunidades, as colunas identificam os grupos ou funções organizacionais. Surpreendentemente, essa matriz é geralmente bastante densa porque muitos grupos querem acesso às mesmas métricas de desempenho principais.

**Priorizando Requisitos**

O documento consolidado de conclusões serve como base para apresentações de volta à alta administração e outros participantes de requisitos. Inegavelmente, você descobriu mais do que pode ser tratado em uma única iteração, então você precisa priorizar. Conforme discutido com o escopo do projeto, não tome essa decisão em um vácuo; você precisa aproveitar (ou fomentar) sua parceria com a comunidade de negócios para estabelecer prioridades apropriadas.

A apresentação de encerramento dos requisitos é posicionada como uma revisão de conclusões e reunião de priorização. Os participantes incluem representantes seniores de negócios (que, idealmente, participaram das entrevistas), bem como o gerente do DW/BI e outros executivos de TI. A sessão começa com uma visão geral de cada processo de negócios identificado. Você quer que todos na sala tenham uma compreensão comum das oportunidades. Também reveja a matriz de oportunidade/partes interessadas, bem como uma matriz de barramento simplificada.

Depois que as conclusões forem apresentadas, é hora de priorizar usando a grade de priorização, ilustrada na Figura 17-2. O eixo vertical da grade refere-se ao impacto potencial ou valor para o negócio. O eixo horizontal transmite viabilidade. Cada um dos temas do processo de negócios das conclusões é colocado na grade com base no acordo composto dos representantes sobre impacto e viabilidade. É visualmente óbvio por onde você deve começar; projetos que justificam atenção imediata estão localizados no canto superior direito porque são projetos de alto impacto, bem como altamente viáveis. Projetos na célula inferior esquerda devem ser evitados como a peste; são missões impossíveis que pouco fazem pelo negócio. Da mesma forma, projetos na célula inferior direita não justificam atenção de curto prazo, embora as equipes de projeto às vezes se gravitem para cá porque esses projetos são viáveis, mas não muito cruciais. Finalmente, projetos na célula superior esquerda representam oportunidades significativas. Esses projetos têm grande potencial de retorno para o negócio, mas atualmente são inviáveis. Enquanto a equipe do projeto DW/BI se concentra em projetos no canto superior direito sombreado, outras equipes de TI devem abordar as limitações de viabilidade atuais daqueles no canto superior esquerdo.



**Tecnologia no Ciclo de Vida**

No roadmap do Ciclo de Vida Kimball na Figura 17-1, a definição de requisitos de negócios é seguida imediatamente por três trilhas concorrentes focadas em tecnologia, dados e aplicativos de BI, respectivamente. Nas próximas seções, vamos nos concentrar na trilha de tecnologia.

**Design da Arquitetura Técnica**

Assim como uma planta para uma nova casa, a arquitetura técnica é a planta para os serviços técnicos e infraestrutura do ambiente DW/BI. Assim como a arquitetura de barramento do data warehouse empresarial apresentada no Capítulo 4 suporta a integração de dados, o plano de arquitetura é uma estrutura organizadora para suportar a integração de tecnologias e aplicativos.

Assim como as plantas residenciais, a arquitetura técnica consiste em uma série de modelos que revelam maiores detalhes sobre cada componente principal. Em ambas as situações, a arquitetura permite identificar problemas no papel (como ter a máquina de lavar louça muito longe da pia) e minimizar surpresas no meio do projeto. Ela suporta a coordenação de esforços paralelos enquanto acelera o desenvolvimento por meio do reuso de componentes modulares. A arquitetura identifica componentes necessários imediatamente versus aqueles que serão incorporados em uma data posterior (como a varanda e a tela de proteção). Mais importante, a arquitetura serve como uma ferramenta de comunicação. Plantas de construção de casas permitem que o arquiteto, o empreiteiro geral, os subcontratados e o proprietário comuniquem-se a partir de um documento comum. Da mesma forma, a arquitetura técnica do DW/BI apoia a comunicação sobre um conjunto consistente de requisitos técnicos dentro da equipe, para a gestão e para fornecedores.

No Capítulo 1, discutimos vários componentes principais da arquitetura, incluindo serviços de ETL e BI. Nesta seção, focamos no processo de criação do design da arquitetura.

Equipes DW/BI geralmente abordam o processo de design da arquitetura de extremos opostos do espectro. Algumas equipes simplesmente não entendem os benefícios de uma arquitetura e sentem que o tema e as tarefas são nebulosos demais. Elas estão tão focadas na entrega que a arquitetura parece uma distração e um impedimento ao progresso, então optam por ignorar o design da arquitetura. Em vez disso, juntam os componentes técnicos necessários para a primeira iteração com "gambiarra", mas a integração e as interfaces ficam sobrecarregadas à medida que mais dados, mais usuários ou mais funcionalidades são adicionados. Eventualmente, essas equipes frequentemente acabam reconstruindo porque a estrutura não arquitetada não suportava as tensões. No outro extremo, algumas equipes querem investir dois anos projetando a arquitetura enquanto esquecem que o principal objetivo de um ambiente DW/BI é resolver problemas de negócios, não abordar qualquer desafio técnico plausível (ou não tão plausível).

Nenhum dos extremos é saudável; a resposta mais adequada está em algum lugar no meio. Identificamos o seguinte processo de oito etapas para ajudar a navegar nessas águas de design arquitetônico. Todo sistema DW/BI tem uma arquitetura técnica; a questão é se ela é planejada e explícita ou apenas implícita.

**Estabeleça uma Força-Tarefa de Arquitetura**

É útil criar uma pequena força-tarefa de duas a três pessoas focadas no design da arquitetura. Normalmente, é o arquiteto técnico, juntamente com o arquiteto/designer de ETL e o arquiteto/designer de aplicativos de BI, que garantem a representação tanto do back room quanto do front room.

**Coletar Requisitos Relacionados à Arquitetura**

Conforme ilustrado na Figura 17-1, definir a arquitetura técnica não é a primeira caixa no diagrama do Ciclo de Vida. A arquitetura é criada para suportar as necessidades de negócios; não é uma desculpa para comprar os produtos mais recentes e melhores. Consequentemente, a entrada chave no processo de design deve vir da definição de requisitos de negócios. No entanto, você deve ouvir os requisitos de negócios com um filtro ligeiramente diferente para direcionar o design da arquitetura. O foco principal é descobrir as implicações arquitetônicas associadas às necessidades de negócios. Ouça atentamente os requisitos de tempo, disponibilidade e desempenho.

Você também deve conduzir entrevistas adicionais dentro da organização de TI. Estas são sessões focadas em tecnologia para entender os padrões atuais, direções técnicas planejadas e limites inegociáveis. Além disso, você deve descobrir lições aprendidas de projetos anteriores de entrega de informações, bem como a disposição da organização para acomodar mudanças operacionais em nome da iniciativa DW/BI, como identificar transações atualizadas no sistema fonte.

**Documentar Requisitos de Arquitetura**

Depois de aproveitar o processo de requisitos de negócios e conduzir entrevistas suplementares de TI, você precisa documentar suas descobertas. Recomendamos o uso de um formato tabular simplista, apenas listando cada requisito de negócio que impacta a arquitetura, juntamente com uma lista de implicações arquitetônicas. Por exemplo, se houver uma necessidade de entregar dados de desempenho de vendas globais diariamente, as implicações técnicas podem incluir disponibilidade 24/7 em todo o mundo, espelhamento de dados para cargas, metadados robustos para suportar acesso global, largura de banda de rede adequada e potência ETL suficiente para lidar com a integração complexa de dados operacionais.

**Criar o Modelo de Arquitetura**

Após documentar os requisitos de arquitetura, você deve começar a formular modelos para suportar as necessidades identificadas. Neste ponto, a equipe de arquitetura muitas vezes se isola em uma sala de conferências por vários dias de pensamento intenso. Os requisitos de arquitetura são agrupados em componentes principais, como ETL, BI, metadados e infraestrutura. A partir daí, a equipe elabora e refina o modelo arquitetônico de alto nível. Este desenho é semelhante à página de elevação frontal em plantas residenciais. Ele ilustra como a arquitetura ficará vista de fora, mas pode ser perigosamente simplista porque detalhes significativos estão embutidos nas páginas que se seguem.

**Determinar Fases de Implementação da Arquitetura**

Como a casa dos sonhos do proprietário, você provavelmente não pode implementar todos os aspectos da arquitetura técnica de uma vez. Alguns são capacidades obrigatórias inegociáveis, enquanto outros são itens desejáveis. Novamente, consulte os requisitos de negócios para estabelecer prioridades de arquitetura porque você deve fornecer minimamente os elementos arquitetônicos necessários para entregar o projeto inicial.

**Projetar e Especificar os Subsistemas**

Uma grande porcentagem da funcionalidade necessária provavelmente será atendida pelas ofertas padrão dos principais fornecedores de ferramentas, mas sempre há alguns subsistemas que podem não ser encontrados em produtos prontos para uso. Você deve definir esses subsistemas em detalhes suficientes para que alguém possa construí-los para você ou para que você possa avaliar produtos de acordo com suas necessidades.

**Criar o Plano de Arquitetura**

A arquitetura técnica precisa ser documentada, incluindo as fases de implementação planejadas, para aqueles que não estavam isolados na sala de conferências. O documento do plano de arquitetura técnica deve incluir detalhes adequados para que profissionais qualificados possam proceder com a construção da estrutura, assim como carpinteiros constroem uma casa com base na planta. No entanto, ele geralmente não faz referência a produtos específicos, exceto aqueles já existentes.

**Revisar e Finalizar a Arquitetura Técnica**

Eventualmente, completamos o círculo com o processo de design da arquitetura. A força-tarefa de arquitetura precisa comunicar o plano de arquitetura em diferentes níveis de detalhe para a equipe do projeto, colegas de TI e líderes de negócios. Após a revisão, a documentação deve ser atualizada e usada imediatamente no processo de seleção de produtos.

**Seleção e Instalação de Produtos**

Em muitos aspectos, o plano de arquitetura é semelhante a uma lista de compras para selecionar produtos que se encaixam na estrutura do plano. As seguintes seis tarefas associadas à seleção de produtos DW/BI são bastante semelhantes a qualquer seleção de tecnologia.

**Entender o Processo de Compra Corporativa**

O primeiro passo antes de selecionar novos produtos é entender os processos internos de compra de hardware e software.

**Desenvolver uma Matriz de Avaliação de Produtos**

Usando o plano de arquitetura como ponto de partida, deve-se desenvolver uma matriz de avaliação baseada em planilha que identifique os critérios de avaliação, juntamente com fatores de ponderação para indicar importância; quanto mais específicos os critérios, melhor. Se os critérios forem muito vagos ou genéricos, todo fornecedor dirá que pode satisfazer suas necessidades.

**Conduzir Pesquisa de Mercado**

Para se tornar compradores informados ao selecionar produtos, você deve fazer uma pesquisa de mercado para entender melhor os participantes e suas ofertas. Um pedido de proposta (RFP) é uma ferramenta clássica de avaliação de produtos. Embora algumas organizações não tenham escolha quanto ao seu uso, você deve evitar essa técnica, se possível. Construir o RFP e avaliar as respostas é extremamente demorado para a equipe. Enquanto isso, os fornecedores são motivados a responder às perguntas de maneira mais positiva, então a avaliação das respostas é frequentemente mais um concurso de beleza. No final, o valor do esforço pode não justificar o investimento.

**Avaliar uma Lista Reduzida de Opções**

Apesar da abundância de produtos disponíveis no mercado, geralmente apenas um pequeno número de fornecedores pode atender aos requisitos de funcionalidade e técnicos. Comparando as pontuações preliminares da matriz de avaliação, você pode se concentrar em uma lista restrita de fornecedores e desqualificar o restante. Depois de lidar com um número limitado de fornecedores, você pode começar as avaliações detalhadas. Representantes de negócios devem estar envolvidos nesse processo se você estiver avaliando ferramentas de BI. Como avaliadores, você deve conduzir o processo em vez de permitir que os fornecedores o conduzam, compartilhando informações relevantes do plano de arquitetura, para que as sessões se concentrem em suas necessidades, e não em recursos adicionais dos produtos. Certifique-se de falar com referências de fornecedores, tanto as fornecidas formalmente quanto as obtidas de sua rede informal.

**Se Necessário, Conduzir um Protótipo**

Depois de realizar as avaliações detalhadas, às vezes um vencedor claro emerge, frequentemente baseado na experiência anterior da equipe ou em relacionamentos. Em outros casos, o líder surge devido a compromissos corporativos existentes, como licenças de site ou compras de hardware legadas. Em qualquer situação, quando um único candidato emerge como o vencedor, você pode ignorar a etapa do protótipo (e o investimento associado tanto em tempo quanto em dinheiro). Se nenhum fornecedor for o vencedor aparente, você deve conduzir um protótipo com no máximo dois produtos. Novamente, assuma o controle do processo desenvolvendo um estudo de caso de negócios limitado, mas realista.

**Selecionar Produto, Instalar em Período de Teste e Negociar**

É hora de selecionar um produto. Em vez de assinar imediatamente na linha pontilhada, preserve seu poder de negociação fazendo um compromisso privado, não público, com um único fornecedor. Em vez de informar ao fornecedor que você está completamente convencido, embarque em um período de teste onde você tenha a oportunidade de colocar o produto em uso real no seu ambiente. Leva-se uma energia significativa para instalar um produto, ser treinado e começar a usá-lo, então você deve seguir esse caminho apenas com o fornecedor que pretende comprar; um teste não deve ser perseguido como outro exercício de avaliação. À medida que o teste se aproxima do fim, você tem a oportunidade de negociar uma compra que seja benéfica para todas as partes envolvidas.

**Trilha de Dados no Ciclo de Vida**

No diagrama do Ciclo de Vida Kimball na Figura 17-1, a trilha do meio que segue a definição de requisitos de negócios se concentra em dados. Voltamos nossa atenção nessa direção nas próximas seções.

**Modelagem Dimensional**

Dado o foco deste livro nos primeiros 16 capítulos, não vamos gastar tempo discutindo técnicas de modelagem dimensional aqui. O próximo capítulo fornece recomendações detalhadas sobre os participantes, processos e entregáveis envolvendo nossa abordagem de workshop iterativo para projetar modelos dimensionais em colaboração com os usuários de negócios. É leitura obrigatória para qualquer pessoa envolvida na atividade de modelagem.

**Design Físico**

Os modelos dimensionais desenvolvidos e documentados por meio de um mapeamento preliminar de origem para destino precisam ser traduzidos em um banco de dados físico. Com a modelagem dimensional, os designs lógico e físico têm uma estreita semelhança; você não quer que o administrador de banco de dados converta seu lindo esquema dimensional em uma estrutura normalizada durante o processo de design físico.

Os detalhes da implementação física do banco de dados variam amplamente por plataforma e projeto. Além disso, hardware, software e ferramentas estão evoluindo rapidamente, então as seguintes atividades e considerações de design físico apenas arranham a superfície.

**Desenvolver Padrões de Nomeação e Banco de Dados**

Nomes de tabelas e colunas são elementos-chave da experiência dos usuários, tanto para navegar no modelo de dados quanto para visualizar aplicativos de BI, então eles devem ser significativos para os negócios. Você também deve estabelecer padrões em torno das declarações de chaves e a permissibilidade de valores nulos.

**Desenvolver Modelo Físico do Banco de Dados**

Este modelo deve ser inicialmente construído no servidor de desenvolvimento onde será usado pela equipe de desenvolvimento de ETL. Há vários conjuntos adicionais de tabelas que precisam ser projetados e implantados como parte do sistema DW/BI, incluindo tabelas de estágio para suportar o sistema ETL, tabelas de auditoria para processamento de ETL e qualidade de dados, e estruturas para suportar acesso seguro a um subconjunto do data warehouse.

**Desenvolver Plano Inicial de Índices**

Além de entender como o otimizador de consultas do banco de dados relacional e os índices funcionam, o administrador do banco de dados também precisa estar ciente de que os requisitos de DW/BI diferem significativamente dos requisitos de OLTP. Como as tabelas de dimensão têm uma chave primária de coluna única, você terá um índice único nessa chave. Se os índices bitmapped estiverem disponíveis, você normalmente adiciona índices bitmapped de coluna única aos atributos de dimensão usados comumente para filtragem e agrupamento, especialmente aqueles atributos que serão restritos conjuntamente; caso contrário, você deve avaliar a utilidade dos índices B-tree nesses atributos. Da mesma forma, o primeiro índice da tabela de fatos será tipicamente um índice B-tree ou clusterizado na chave primária; colocar a chave estrangeira de data na posição de liderança do índice acelera tanto as cargas de dados quanto as consultas porque a data é frequentemente restringida. Se o SGBD suportar índices bitmapped de alta cardinalidade, estes podem ser uma boa escolha para chaves estrangeiras individuais nas tabelas de fatos porque são mais agnósticos do que índices clusterizados quando o usuário restringe dimensões de maneiras inesperadas. A determinação de outros índices na tabela de fatos depende das opções de índice e estratégias de otimização dentro da plataforma. Embora motores de banco de dados OLAP também usem índices e tenham um otimizador de consultas, ao contrário do mundo relacional, o administrador do banco de dados tem pouco controle nesses ambientes.

**Design de Agregações, Incluindo Banco de Dados OLAP**

Ao contrário da crença popular, adicionar mais hardware não é necessariamente a melhor arma no arsenal de ajuste de desempenho; aproveitar tabelas de agregação é uma alternativa muito mais econômica. Quer use tecnologia OLAP ou tabelas de agregação relacionais, agregações precisam ser projetadas no ambiente DW/BI, como exploraremos mais nos Capítulos 19: Subsystemas e Técnicas de ETL, e Capítulo 20. Quando métricas de desempenho são agregadas, você elimina dimensões ou associa as métricas a uma dimensão de rollup reduzida que conforma com a dimensão base atômica. Porque você não pode construir, armazenar e administrar todas as agregações teóricas, dois fatores principais precisam ser avaliados. Primeiro, pense sobre os padrões de acesso dos usuários de negócios derivados das descobertas de requisitos, bem como das entradas obtidas monitorando padrões de uso reais. Segundo, avalie a distribuição estatística dos dados para identificar pontos de agregação que entregam um bom custo-benefício.

**Finalizar Detalhes de Armazenamento Físico**

Isso inclui as estruturas de armazenamento básicas de blocos, arquivos, discos, partições e espaços de tabelas ou bancos de dados. Grandes tabelas de fatos são tipicamente particionadas por data de atividade, com dados segmentados por mês em partições separadas enquanto aparecem aos usuários como uma única tabela. Particionar por data oferece vantagens de carregamento de dados, manutenção e desempenho de consultas.

As estratégias de agregação, indexação e outras de ajuste de desempenho evoluirão à medida que os padrões de uso reais forem melhor compreendidos, então esteja preparado para as inevitáveis modificações contínuas. No entanto, você deve entregar dados apropriadamente indexados e agregados com o lançamento inicial para garantir que o ambiente DW/BI ofereça desempenho de consulta razoável desde o início.

**Design e Desenvolvimento de ETL**

A trilha de dados do Ciclo de Vida termina com o design e desenvolvimento do sistema ETL. O Capítulo 19 descreve os fatores, apresentados como 34 subsistemas, que devem ser considerados durante o design. O Capítulo 20 então fornece orientação mais granular sobre o processo de design e desenvolvimento do sistema ETL e as tarefas associadas. Fique atento para mais detalhes sobre ETL.

**Trilha de Aplicativos de BI no Ciclo de Vida**

O conjunto final de atividades paralelas que seguem a definição de requisitos de negócios na Figura 17-1 é a trilha de aplicativos de BI onde você projeta e desenvolve os aplicativos que atendem a uma parte dos requisitos analíticos dos usuários. Como um desenvolvedor de aplicativos de BI disse uma vez, "Lembre-se, esta é a parte divertida!" Você está finalmente usando o investimento em tecnologia e dados para ajudar os usuários de negócios a tomar melhores decisões.

Embora alguns possam sentir que o data warehouse deve ser um ambiente de consulta totalmente ad hoc e de autoatendimento, entregar aplicativos de BI orientados por parâmetros satisfará uma grande porcentagem das necessidades da comunidade de negócios. Para muitos usuários de negócios, "ad hoc" implica a capacidade de alterar os parâmetros em um relatório para criar sua versão personalizada. Não faz sentido fazer com que cada usuário comece do zero.

Construir um conjunto de aplicativos de BI estabelece uma estrutura analítica consistente para a organização, em vez de permitir que cada planilha conte uma história ligeiramente diferente. Aplicativos de BI também servem para capturar a expertise analítica da organização, desde monitorar o desempenho até identificar exceções, determinar fatores causais e modelar respostas alternativas; essa encapsulação fornece um impulso inicial para os menos inclinados analiticamente.

**Especificação de Aplicativos de BI**

Após a definição de requisitos de negócios, você precisa revisar as conclusões e relatórios de amostra coletados para identificar um conjunto inicial de aproximadamente 10 a 15 relatórios de BI e aplicativos analíticos. Você quer restringir o foco inicial às capacidades mais críticas para gerenciar expectativas e garantir a entrega pontual. A entrada da comunidade de negócios será crítica para esse processo de priorização. Embora 15 aplicativos possam não parecer muito, inúmeras análises podem ser criadas a partir de um único template apenas alterando variáveis.

Antes de começar a projetar os aplicativos iniciais, é útil estabelecer padrões, como menus suspensos comuns e uma aparência consistente de saída. Usando esses padrões, você especifica cada template de aplicativo e captura informações suficientes sobre o layout, variáveis de entrada, cálculos e quebras, para que tanto o desenvolvedor de aplicativos quanto os representantes de negócios compartilhem um entendimento comum.

Durante a atividade de especificação do aplicativo de BI, você também deve considerar a organização dos aplicativos. Você precisa identificar caminhos de navegação estruturados para acessar os aplicativos, refletindo a forma como os usuários pensam sobre seus negócios. Aproveitar portais de informação ou dashboards personalizáveis são as estratégias dominantes para disseminar o acesso.

**Desenvolvimento de Aplicativos de BI**

Quando você entra na fase de desenvolvimento dos aplicativos de BI, novamente precisa se concentrar em padrões; convenções de nomenclatura, cálculos, bibliotecas e padrões de codificação devem ser estabelecidos para minimizar retrabalho futuro. A atividade de desenvolvimento de aplicativos pode começar quando o design do banco de dados estiver completo, as ferramentas de BI e metadados estiverem instaladas e um subconjunto de dados históricos tiver sido carregado. As especificações do template de aplicativos de BI devem ser revisitadas para levar em conta as inevitáveis mudanças no modelo desde que as especificações foram concluídas.

Cada ferramenta de BI tem truques específicos do produto que podem fazer com que ela execute tarefas complexas. Em vez de tentar aprender as técnicas por tentativa e erro, sugerimos investir em educação específica para a ferramenta ou recursos suplementares para a equipe de desenvolvimento.

Enquanto os aplicativos de BI estão sendo desenvolvidos, vários benefícios auxiliares resultam. Desenvolvedores de aplicativos de BI, armados com uma ferramenta de acesso robusta, encontrarão rapidamente problemas na qualidade dos dados apesar da garantia de qualidade realizada pelo aplicativo de ETL. Este é um dos motivos pelos quais preferimos iniciar a atividade de desenvolvimento de aplicativos de BI antes da suposta conclusão do sistema ETL. Os desenvolvedores também serão os primeiros a testar realisticamente os tempos de resposta das consultas. Agora é o momento de revisar as estratégias preliminares de ajuste de desempenho.

As atividades de garantia de qualidade dos aplicativos de BI não podem ser concluídas até que os dados estejam estabilizados. Você deve garantir que haja tempo adequado no cronograma além do corte final do ETL para permitir um encerramento ordenado das tarefas de desenvolvimento dos aplicativos de BI.

**Atividades de Encerramento do Ciclo de Vida**

As seções seguintes fornecem recomendações para garantir que seu projeto chegue a uma conclusão ordenada, garantindo que você esteja pronto para a expansão futura.

**Implantação**

As trilhas de tecnologia, dados e aplicativos de BI convergem na implantação. Infelizmente, essa convergência não acontece naturalmente, mas requer um planejamento substancial. Talvez mais importante, a implantação bem-sucedida exige coragem e força de vontade para avaliar honestamente a prontidão do projeto para implantar. A implantação é semelhante a servir uma grande refeição de feriado para amigos e parentes. Pode ser difícil prever exatamente quanto tempo levará para cozinhar o prato principal da refeição. Claro, se o prato principal não estiver pronto, o cozinheiro é forçado a desacelerar os acompanhamentos para compensar o atraso antes de chamar todos para a mesa.

No caso da implantação do DW/BI, os dados são o prato principal. "Cozinhar" os dados na cozinha do ETL é a tarefa mais imprevisível. Infelizmente, mesmo se os dados não estiverem completamente preparados, muitas vezes você ainda prossegue com a implantação do DW/BI porque informou aos convidados do data warehouse que eles seriam servidos em uma data e hora específicas. Porque você não está disposto a desacelerar o ritmo da implantação, você marcha nos escritórios deles com dados mal cozidos. Não é de se admirar que os usuários às vezes evitem voltar para uma segunda porção.

Embora os testes tenham indubitavelmente ocorrido durante as tarefas de desenvolvimento do DW/BI, você precisa realizar testes de sistema de ponta a ponta, incluindo garantia de qualidade dos dados, processamento de operações, desempenho e usabilidade. Além de avaliar criticamente a prontidão dos entregáveis do DW/BI, você também precisa empacotá-los com educação e suporte para a implantação. Porque a comunidade de usuários deve adotar o sistema DW/BI para que ele seja considerado bem-sucedido, a educação é crítica. A estratégia de suporte ao DW/BI depende de uma combinação das expectativas da gestão e das realidades dos entregáveis. O suporte é frequentemente organizado em uma estrutura escalonada. O primeiro nível é o suporte por site e autoatendimento; o segundo nível é fornecido pelos usuários avançados que residem na área de negócios; o suporte centralizado da equipe de DW/BI fornece a linha final de defesa.

**Manutenção e Crescimento**

Você passou pela implantação, então agora está pronto para relaxar. Não tão rápido! Seu trabalho está longe de estar completo após a implantação. Você precisa continuar a gerenciar o ambiente existente investindo recursos nas seguintes áreas:

**Suporte:** O suporte ao usuário é crucial imediatamente após a implantação para garantir que a comunidade de negócios fique engajada. Você não pode sentar em sua baia e assumir que a falta de notícias da comunidade de negócios é um bom sinal. Se você não está ouvindo deles, provavelmente ninguém está usando o sistema DW/BI. Realoque-se (pelo menos temporariamente) para a comunidade de negócios para que os usuários tenham acesso fácil aos recursos de suporte. Se problemas com os dados ou aplicativos de BI forem descobertos, seja honesto com os negócios para construir credibilidade enquanto toma ações imediatas para corrigir os problemas. Se o entregável do DW/BI não for de alta qualidade, as demandas de suporte não antecipadas para reconciliação de dados e retrabalho de aplicativos podem ser esmagadoras.

**Educação:** Você deve fornecer um programa contínuo de educação para o sistema DW/BI. O currículo deve incluir cursos formais de reciclagem e avançados, bem como repetir cursos introdutórios. Educação mais informal pode ser oferecida aos desenvolvedores e usuários avançados para incentivar a troca de ideias.

**Suporte técnico:** O sistema DW/BI precisa ser tratado como um ambiente de produção com acordos de nível de serviço. Claro, o suporte técnico deve monitorar proativamente as tendências de desempenho e capacidade do sistema. Você não quer depender da comunidade de negócios para lhe dizer que o desempenho diminuiu.

**Suporte ao programa:** O programa DW/BI continua além da implementação de uma única fase. Você deve monitorar de perto e depois comercializar seu sucesso. A comunicação com as várias partes interessadas do DW/BI deve continuar. Você também deve garantir que as implementações existentes continuem a atender às necessidades dos negócios. Revisões contínuas de pontos de verificação são uma ferramenta chave para avaliar e identificar oportunidades de melhoria.

Se você fez seu trabalho corretamente, inevitavelmente haverá demanda por crescimento, seja para novos usuários, novos dados, novos aplicativos de BI ou grandes melhorias nos entregáveis existentes. Ao contrário das iniciativas tradicionais de desenvolvimento de sistemas, a mudança no DW/BI deve ser vista como um sinal de sucesso, não de fracasso. Como aconselhamos anteriormente ao discutir o escopo do projeto, a equipe de DW/BI não deve tomar decisões sobre essas opções de crescimento em um vácuo; os negócios precisam estar envolvidos no processo de priorização. Este é um bom momento para aproveitar a grade de priorização ilustrada na Figura 17-2. Se você ainda não o fez, um comitê de patrocínio executivo de negócios deve ser estabelecido para definir as prioridades do DW/BI que se alinham com os objetivos gerais da organização. Depois que novas prioridades forem identificadas, você volta ao início do Ciclo de Vida e faz tudo de novo, aproveitando e construindo sobre as fundações técnicas, de dados e de aplicativos de BI que já foram estabelecidas, enquanto volta sua atenção para os novos requisitos.

**Armadilhas Comuns a Evitar**

Embora possamos fornecer recomendações positivas sobre data warehousing e business intelligence, alguns leitores se relacionam melhor com uma lista de armadilhas comuns. Aqui está nossa lista dos 10 erros mais comuns a evitar ao construir um sistema DW/BI. Estes são todos erros bastante letais — um só pode ser suficiente para derrubar a iniciativa:

* Ficar excessivamente encantado com tecnologia e dados, em vez de focar nos requisitos e objetivos dos negócios.
* Não abraçar ou recrutar um visionário sênior influente, acessível e razoável como patrocinador do negócio para o esforço do DW/BI.
* Lidar com um projeto galáctico de vários anos, em vez de buscar esforços de desenvolvimento iterativo mais gerenciáveis, embora ainda atraentes.
* Alocar energia para construir uma estrutura de dados normalizada, mas esgotar o orçamento antes de construir uma área de apresentação viável baseada em modelos dimensionais.
* Prestar mais atenção ao desempenho operacional de back room e facilidade de desenvolvimento do que ao desempenho de consulta de front room e facilidade de uso.
* Tornar os dados supostamente consultáveis na área de apresentação excessivamente complexos. Designers de banco de dados que preferem uma apresentação mais complexa deveriam passar um ano apoiando usuários de negócios; eles desenvolveriam uma apreciação muito melhor pela necessidade de buscar soluções mais simples.
* Popular modelos dimensionais de forma isolada sem considerar uma arquitetura de dados que os conecte usando dimensões conformes compartilhadas.
* Carregar apenas dados resumidos nas estruturas dimensionais da área de apresentação.
* Presumir que os negócios, seus requisitos e análises, os dados subjacentes e a tecnologia de suporte são estáticos.
* Negligenciar reconhecer que o sucesso do DW/BI está diretamente ligado à aceitação pelos negócios. Se os usuários não aceitaram o sistema DW/BI como uma base para a tomada de decisões aprimorada, seus esforços foram exercícios em vão.

Este capítulo forneceu uma visão geral do Ciclo de Vida Kimball para projetos DW/BI.Tocamos nos processos principais e melhores práticas. Embora cada projeto seja um pouco diferente do próximo, todos eles exigem atenção às principais tarefas discutidas para garantir uma iniciativa bem-sucedida.

O próximo capítulo fornece uma cobertura muito mais detalhada da abordagem de workshop colaborativo do Ciclo de Vida Kimball para projetar iterativamente modelos dimensionais com representantes de negócios. Os Capítulos 19 e 20 mergulham em considerações de design do sistema ETL e processos de desenvolvimento recomendados.

**Capítulo 18**

**Visão Geral do Ciclo de Vida do Kimball DW/BI**

**Processo de Modelagem Dimensional e Tarefas**

Descrevemos inúmeros padrões de modelagem dimensional nos Capítulos 1 a 16 deste livro. Agora é hora de voltar sua atenção para as tarefas e táticas do processo de modelagem dimensional. Este capítulo, condensado do conteúdo do livro The Data Warehouse Lifecycle Toolkit, Second Edition (Wiley, 2008), começa com uma discussão prática sobre atividades preliminares de preparação, como identificar os participantes (incluindo representantes de negócios) e organizar a logística. A equipe de modelagem desenvolve um diagrama inicial de modelo de alto nível, seguido por um desenvolvimento detalhado do modelo, revisão e validação iterativos. Durante todo o processo, você está reconfirmando sua compreensão dos requisitos de negócios.

**O Capítulo 18 revisa os seguintes conceitos:**

1. Visão geral do processo de modelagem dimensional
2. Recomendações táticas para as tarefas de modelagem
3. Principais entregáveis da modelagem
4. Visão Geral do Processo de Modelagem

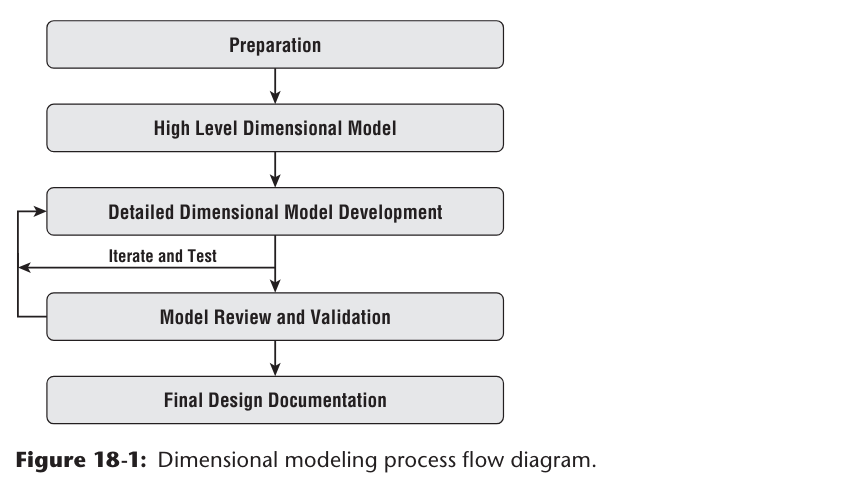
Antes de iniciar o esforço de design da modelagem dimensional, você deve envolver as pessoas certas. Notavelmente, recomendamos fortemente a participação de representantes de negócios durante as sessões de modelagem. Seu envolvimento e colaboração aumentam significativamente a probabilidade de que o modelo resultante atenda às necessidades dos negócios. Da mesma forma, os responsáveis pelos dados da organização devem participar, especialmente quando você estiver discutindo os dados pelos quais eles são responsáveis por governar.

Criar um modelo dimensional é um processo altamente iterativo e dinâmico. Após alguns passos de preparação, o esforço de design começa com um modelo gráfico inicial derivado da matriz de barramento, identificando o escopo do design e esclarecendo o grão das tabelas de fatos propostas e as dimensões associadas.

Após concluir o modelo de alto nível, a equipe de design mergulha nas tabelas de dimensões com definições de atributos, valores de domínio, fontes, relacionamentos, preocupações de qualidade de dados e transformações. Depois de identificar as dimensões, as tabelas de fatos são modeladas. A última fase do processo envolve revisar e validar o modelo com as partes interessadas, especialmente representantes de negócios. Os objetivos principais são criar um modelo que atenda aos requisitos de negócios, verificar se os dados estão disponíveis para preencher o modelo e fornecer à equipe de ETL um mapeamento de origem para destino sólido.

Os modelos dimensionais se desdobram por meio de uma série de sessões de design, com cada passagem resultando em um design mais detalhado e robusto que foi repetidamente testado contra as necessidades do negócio. O processo é concluído quando o modelo atende claramente aos requisitos de negócios. Um design típico requer de três a quatro semanas para um modelo dimensional de um único processo de negócios, mas o tempo necessário pode variar dependendo da experiência da equipe, da disponibilidade de requisitos detalhados de negócios, do envolvimento de representantes de negócios ou administradores de dados autorizados a alcançar o consenso organizacional, da complexidade dos dados de origem e da capacidade de aproveitar as dimensões conformadas existentes.

A Figura 18-1 mostra o fluxo do processo de modelagem dimensional. As principais entradas para o processo de modelagem dimensional são a matriz de barramento preliminar e os requisitos detalhados de negócios. Os principais entregáveis do processo de modelagem são o modelo dimensional de alto nível, os designs detalhados das tabelas de dimensões e de fatos, e o log de questões. Embora o gráfico retrate uma progressão linear, o processo é bastante iterativo. Você fará várias passagens pelo modelo dimensional, começando em um nível alto e aprofundando-se em cada tabela e coluna, preenchendo as lacunas, adicionando mais detalhes e alterando o design com base em novas informações.

  
 Se um especialista externo for contratado para ajudar a guiar o esforço de modelagem dimensional, insista para que ele facilite o processo com a equipe, em vez de desaparecer por algumas semanas e retornar com um design concluído. Isso garante que toda a equipe compreenda o design e os trade-offs associados. Também proporciona uma oportunidade de aprendizado, para que a equipe possa levar o modelo adiante e enfrentar o próximo modelo de forma independente.

**Organize-se**

Antes de começar a modelar, você deve se preparar adequadamente para o processo de modelagem dimensional. Além de envolver os recursos certos, há também considerações logísticas básicas para garantir um esforço de design produtivo.

**Identifique os Participantes, Especialmente Representantes de Negócios**

Os melhores modelos dimensionais resultam de um esforço de equipe colaborativo. Nenhum indivíduo sozinho é provável que tenha o conhecimento detalhado dos requisitos de negócios e das peculiaridades dos sistemas de origem para criar o modelo de forma eficaz. Embora o modelador de dados facilite o processo e tenha responsabilidade principal pelos entregáveis, acreditamos que é de importância crítica envolver especialistas no assunto dos negócios para colaborar ativamente; seus insights são inestimáveis, especialmente porque muitas vezes são as pessoas que historicamente descobriram como obter dados dos sistemas de origem e transformá-los em informações analíticas valiosas. Embora envolver mais pessoas nas atividades de design aumente o risco de desacelerar o processo, a riqueza e a completude aprimoradas do design justificam a sobrecarga adicional.

É sempre útil ter alguém com conhecimento aguçado das realidades do sistema de origem envolvido. Você também pode incluir alguns representantes da equipe de DBA física e ETL para que possam aprender com os insights descobertos durante o esforço de modelagem e resistir às tentações de aplicar conceitos de terceira forma normal (3NF) ou deferir complexidades para as aplicações de BI em um esforço para simplificar o processamento de ETL. Lembre-se de que o objetivo é trocar a complexidade do processamento de ETL por simplicidade e previsibilidade na camada de apresentação de BI.

Antes de entrar no processo de modelagem, você deve considerar a gestão contínua do ambiente DW/BI. Se a organização tiver uma iniciativa ativa de governança e gestão de dados, é hora de aproveitar essa função. Se não houver um programa de gestão pré-existente, é hora de iniciá-lo. Um esforço de DW/BI empresarial comprometido com a modelagem dimensional também deve estar comprometido com uma estratégia de dimensão conformada para garantir consistência entre os processos de negócios. Um programa ativo de gestão de dados ajuda a organização a alcançar sua estratégia de dimensão conformada. Concordar com dimensões conformadas em uma grande empresa pode ser um desafio; a dificuldade geralmente é menos um problema técnico e mais um desafio de comunicação e construção de consenso organizacional.

Diferentes grupos na empresa costumam estar comprometidos com suas próprias regras e definições de negócios proprietárias. Os administradores de dados devem trabalhar de perto com os grupos interessados para desenvolver regras e definições de negócios comuns, e depois persuadir a organização a adotar as regras e definições comuns para desenvolver um consenso empresarial. Ao longo dos anos, alguns criticaram o conceito de dimensões conformadas por serem “muito difíceis”. Sim, é difícil fazer com que pessoas de diferentes cantos dos negócios concordem com nomes, definições e valores de atributos comuns, mas esse é o cerne de dados unificados e integrados. Se todos exigirem seus próprios rótulos e regras de negócios, então não há chance de entregar a versão única da verdade prometida pelos sistemas DW/BI. E finalmente, uma das razões pelas quais a abordagem Kimball às vezes é criticada como sendo difícil por pessoas que estão procurando soluções rápidas é porque detalhamos os passos necessários para realmente fazer o trabalho. No Capítulo 19: ETL Subsystems and Techniques, esses detalhes minuciosos são discutidos na cobertura dos subsistemas de ETL 17 e 18.

**Revise os Requisitos de Negócios**

Antes de iniciar a modelagem, a equipe deve se familiarizar com os requisitos de negócios. O primeiro passo é revisar cuidadosamente a documentação de requisitos, como descrevemos no Capítulo 17: Kimball DW/BI Lifecycle Overview. É responsabilidade da equipe de modelagem traduzir os requisitos de negócios em um modelo dimensional flexível que possa suportar uma ampla gama de análises, não apenas relatórios específicos. Alguns designers são tentados a pular a revisão dos requisitos e passar diretamente para o design, mas os modelos resultantes geralmente são impulsionados exclusivamente pelos dados de origem sem considerar o valor agregado exigido pela comunidade de negócios. Ter uma representação adequada de negócios na equipe de modelagem ajuda a evitar essa abordagem orientada por dados.

**Aproveite uma Ferramenta de Modelagem**

Antes de começar as atividades de modelagem, é útil ter algumas ferramentas em mãos. Usar uma planilha como ferramenta de documentação inicial é eficaz porque permite que você faça mudanças rápida e facilmente enquanto itera pelo processo de modelagem. Depois que o modelo começa a se firmar nas etapas posteriores do processo, você pode converter para qualquer ferramenta de modelagem usada em sua organização. A maioria das ferramentas de modelagem são cientes dimensionalmente, com funções para suportar a criação de um modelo dimensional. Quando o design detalhado estiver completo, as ferramentas de modelagem podem ajudar o DBA a projetar o modelo para o banco de dados, incluindo a criação das tabelas, índices, partições, visualizações e outros elementos físicos do banco de dados.

**Aproveite uma Ferramenta de Perfil de Dados**

Durante todo o processo de modelagem, as equipes precisam desenvolver uma compreensão crescente da estrutura, conteúdo, relacionamentos e regras de derivação dos dados de origem. Você precisa verificar se os dados existem em um estado utilizável, ou pelo menos se suas falhas podem ser gerenciadas, e entender o que é necessário para convertê-los no modelo dimensional. A profilagem de dados usa capacidades de consulta para explorar o conteúdo real e os relacionamentos no sistema de origem, em vez de confiar em documentação possivelmente incompleta ou desatualizada. A profilagem de dados pode ser tão simples quanto escrever algumas instruções SQL ou tão sofisticada quanto uma ferramenta de propósito especial. Os principais fornecedores de ETL incluem capacidades de profilagem de dados em seus produtos.

**Aproveite ou Estabeleça Convenções de Nomenclatura**

A questão das convenções de nomenclatura inevitavelmente surge durante a criação do modelo dimensional. Os rótulos do modelo de dados devem ser descritivos e consistentes do ponto de vista dos negócios. Os nomes de tabelas e colunas tornam-se elementos chave da interface da aplicação de BI. Um nome de coluna como "Descrição" pode ser perfeitamente claro no contexto de um modelo de dados, mas não comunica nada no contexto de um relatório. Parte do processo de design de um modelo dimensional é concordar com definições e rótulos comuns. A nomenclatura é complexa porque diferentes grupos de negócios têm significados diferentes para o mesmo nome e nomes diferentes com o mesmo significado. As pessoas relutam em abandonar o familiar e adotar um novo vocabulário. Gastar tempo em convenções de nomenclatura é uma dessas tarefas tediosas que parecem ter pouco retorno, mas que vale a pena a longo prazo.

Grandes organizações frequentemente têm uma função de TI responsável pelas convenções de nomenclatura. Uma abordagem comum é usar um padrão de nomenclatura com três partes: palavra principal, qualificadores (se apropriados) e palavra de classe. Aproveite o trabalho dessa função de TI, entendendo que, às vezes, as convenções de nomenclatura existentes precisam ser estendidas para suportar nomes de tabelas e colunas mais amigáveis aos negócios. Se a organização ainda não tiver um conjunto de convenções de nomenclatura, você deve estabelecê-las durante a modelagem dimensional.

**Coordene Calendários e Instalações**

Por último, mas não menos importante, você precisa agendar as sessões de design nos calendários dos participantes. Em vez de tentar reservar dias inteiros, é mais realista agendar sessões matinais e vespertinas que durem de duas a três horas, três ou quatro dias por semana. Essa abordagem reconhece que os membros da equipe têm outras responsabilidades e permite que eles tentem acompanhar nos horários antes, depois e entre as sessões de design. A equipe de design pode aproveitar o tempo não agendado para pesquisar os dados de origem e confirmar os requisitos, bem como permitir tempo para o modelador de dados atualizar a documentação do design antes de cada sessão.

Como mencionamos anteriormente, o processo de modelagem normalmente leva de três a quatro semanas para um único processo de negócios, como pedidos de vendas, ou um par de processos de negócios intimamente relacionados, como transações de reivindicação de instalações de saúde e profissionais em um conjunto de tabelas de fatos distintas, mas estreitamente alinhadas. Existem vários fatores que impactam a magnitude do esforço. Em última análise, a disponibilidade de dimensões principais previamente existentes permite que o esforço de modelagem se concentre quase exclusivamente nas métricas de desempenho da tabela de fatos, o que reduz significativamente o tempo necessário.

Finalmente, você deve reservar instalações apropriadas. É melhor reservar uma sala de conferências dedicada para a duração do esforço de design - uma tarefa difícil na maioria das organizações onde as instalações de salas de reuniões são sempre escassas. Embora estejamos sonhando, quadros brancos grandes, do chão ao teto, em todas as quatro paredes seriam ótimos também! Além de uma instalação de reuniões, a equipe precisa de alguns suprimentos básicos, como papel de flip chart autoadesivo. Um projetor de laptop é frequentemente útil durante as sessões de design e é absolutamente necessário para as revisões de design.

**Desenhe o Modelo Dimensional**

Conforme delineado no Capítulo 3: Vendas no Varejo, há quatro decisões chave tomadas durante o design de um modelo dimensional:

* Identifique o processo de negócios.
* Declare o grão do processo de negócios.
* Identifique as dimensões.
* Identifique os fatos.

O primeiro passo de identificar o processo de negócios é normalmente determinado na conclusão da coleta de requisitos. A atividade de priorização descrita no Capítulo 17 estabelece qual linha da matriz de barramento (e, portanto, processo de negócios) será modelada. Com isso em mente, a equipe pode prosseguir com as tarefas de design.

O esforço de modelagem normalmente segue a sequência de tarefas e entregáveis ilustrada na Figura 18-1:

* Modelo de alto nível definindo o escopo e a granularidade do modelo
* Design detalhado com atributos e métricas de cada tabela
* Revisão e validação com representantes de TI e negócios
* Finalização da documentação do design

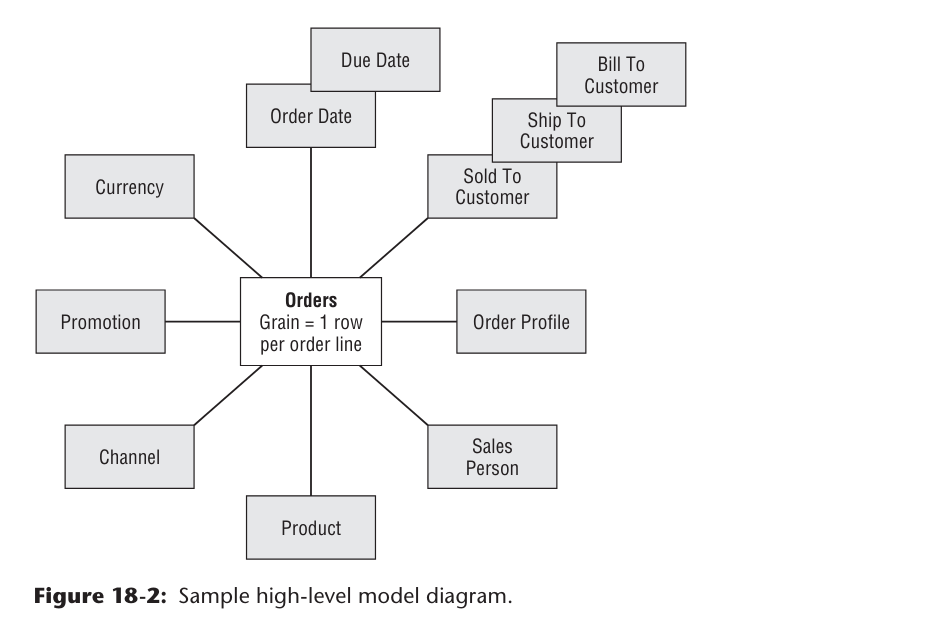
Como qualquer esforço de modelagem de dados, a modelagem dimensional é um processo iterativo. Você trabalhará de volta e meia entre os requisitos de negócios e os detalhes de origem para refinar ainda mais o modelo, alterando o modelo conforme aprende mais.

Esta seção descreve cada uma dessas tarefas principais. Dependendo da experiência da equipe de design e da exposição aos conceitos de modelagem dimensional, você pode começar com uma educação básica em modelagem dimensional antes de iniciar o esforço para garantir que todos estejam na mesma página em relação ao vocabulário dimensional padrão e às melhores práticas.

**Alcance o Consenso no Diagrama de Bolhas de Alto Nível**

A tarefa inicial na sessão de design é criar um diagrama de modelo dimensional de alto nível para o processo de negócios alvo. Criar o primeiro rascunho é relativamente simples porque você começa com a matriz de barramento. Embora um designer experiente possa desenvolver o modelo dimensional de alto nível inicial e apresentá-lo à equipe para revisão, recomendamos contra essa abordagem porque não permite que toda a equipe participe do processo.

O diagrama de alto nível representa graficamente as tabelas de dimensão e de fatos do processo de negócios. Mostrado na Figura 18-2, frequentemente nos referimos a este diagrama como o diagrama de bolhas por razões óbvias. Este modelo gráfico em nível de entidade identifica claramente o grão da tabela de fatos e suas dimensões associadas para um público não técnico. Declarar o grão exige que a equipe de modelagem considere o que é necessário para atender aos requisitos de negócios e o que é possível com base nos dados coletados pelo sistema de origem. O diagrama de bolhas deve estar enraizado nas realidades das fontes de dados físicas disponíveis. Uma única linha da matriz de barramento pode resultar em múltiplos diagramas de bolhas, cada um correspondendo a uma tabela de fatos única com granularidade única.



A maioria das principais dimensões surgirá naturalmente após você determinar o grão. Um dos efeitos poderosos de uma declaração clara do grão da tabela de fatos é que você pode visualizar precisamente a dimensionalidade associada. Escolher as dimensões pode também fazer você repensar a declaração do grão. Se uma dimensão proposta não corresponder ao grão da tabela de fatos, a dimensão deve ser deixada de fora, o grão da tabela de fatos deve ser alterado, ou uma solução de design multivalorada deve ser considerada. A representação gráfica da Figura 18-2 serve a vários propósitos. Facilita a discussão dentro da equipe de design antes de a equipe mergulhar no design detalhado, garantindo que todos estejam na mesma página antes de serem inundados com minúcias. Também é uma introdução útil quando a equipe se comunica com as partes interessadas sobre o projeto, seu escopo e conteúdo dos dados.

Para ajudar na compreensão, é útil manter a consistência nos diagramas de modelo de alto nível para um dado processo de negócios. Embora cada tabela de fatos seja documentada em uma página separada, organizar as dimensões associadas em uma sequência semelhante nos diagramas de bolhas é útil.

**Desenvolva o Modelo Dimensional Detalhado**

Após concluir os designs do diagrama de bolhas de alto nível, é hora de focar nos detalhes. A equipe deve se reunir regularmente para definir o modelo dimensional detalhado, tabela por tabela, coluna por coluna. Os representantes de negócios devem permanecer engajados durante essas sessões interativas; você precisa do feedback deles sobre atributos, filtros, agrupamentos, rótulos e métricas.

É mais eficaz começar com as tabelas de dimensões e depois trabalhar nas tabelas de fatos. Sugerimos iniciar o processo de design detalhado com algumas dimensões simples; a dimensão de data é sempre um ponto de partida favorito. Isso permite que a equipe de modelagem obtenha sucesso inicial, desenvolva uma compreensão do processo de modelagem e aprenda a trabalhar em conjunto como uma equipe.

A modelagem detalhada identifica os atributos interessantes e úteis dentro de cada dimensão e as métricas apropriadas para cada tabela de fatos. Você também deseja capturar as fontes, definições e regras preliminares de negócios que especificam como esses atributos e métricas são populados. Análises contínuas do sistema de origem e profilagem sistemática de dados durante as sessões de design ajudam a equipe a entender melhor as realidades dos dados de origem subjacentes.

**Identifique Dimensões e seus Atributos**

Durante as sessões de design detalhado, são definidas as principais dimensões conformadas. Como o sistema DW/BI é um recurso empresarial, essas definições devem ser aceitáveis em toda a empresa. Os administradores de dados e analistas de negócios são recursos chave para alcançar o consenso organizacional sobre nomes de tabelas e atributos, descrições e definições. A equipe de design pode liderar o processo e aproveitar as convenções de nomenclatura, se disponíveis. Mas, em última análise, é uma tarefa de negócios concordar com definições e nomes de negócios padrão; os nomes das colunas devem fazer sentido para os usuários de negócios. Isso pode levar algum tempo, mas é um investimento que trará grandes retornos para a compreensão e aceitação do modelo dimensional pelos usuários. Não se surpreenda se o comitê de governança precisar se envolver para resolver problemas de definição e nomenclatura de dimensões conformadas.

Neste ponto, a equipe de modelagem muitas vezes também luta com a potencial inclusão de dimensões lixo ou mini-dimensões em um modelo dimensional. Pode não ser aparente que esses padrões mais centrados em desempenho sejam justificados até que a equipe esteja profundamente imersa no design.

**Identifique os Fatos**

Declarar o grão cristaliza a discussão sobre as métricas da tabela de fatos porque os fatos devem ser verdadeiros para o grão. O esforço de profilagem de dados identifica as contagens e montantes gerados pelo evento de medição do sistema de origem. No entanto, as tabelas de fatos não se limitam apenas a esses fatos base. Pode haver métricas adicionais que os negócios desejam analisar e que são derivadas dos fatos base.

**Identifique Técnicas de Dimensão de Mudança Lenta**

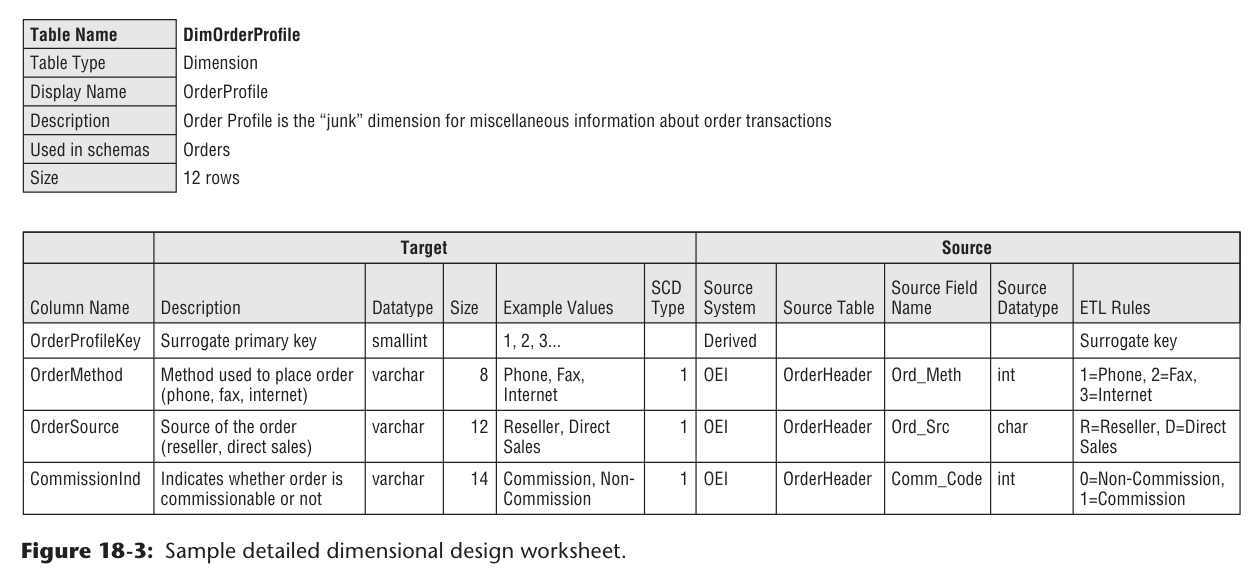
Após as tabelas de dimensões e de fatos do diagrama de modelo de alto nível terem sido inicialmente elaboradas, você então volta para as tabelas de dimensões. Para cada atributo da tabela de dimensões, você define como as mudanças nos dados do sistema de origem serão refletidas na tabela de dimensões. Novamente, a entrada dos administradores de dados de negócios é crítica para estabelecer regras apropriadas. Também é útil perguntar aos especialistas do sistema de origem se eles podem determinar se uma mudança no elemento de dados é devido a uma correção de dados de origem.

**Documente os Designs Detalhados das Tabelas**

Os principais entregáveis da fase de modelagem detalhada são as planilhas de design, como mostrado na Figura 18-3; um modelo digital está disponível em nosso site em www.kimballgroup.com na guia Tools and Utilities para The Data Warehouse Lifecycle Toolkit, Second Edition. As planilhas capturam detalhes para comunicação com as partes interessadas, incluindo outros usuários analíticos de negócios, desenvolvedores de aplicativos de BI e, mais importante, os desenvolvedores de ETL que serão responsáveis por popular o design.

Cada tabela de dimensões e de fatos deve ser documentada em uma planilha separada. No mínimo, as informações de suporte necessárias incluem o nome do atributo/fato, descrição, valores de exemplo e um indicador do tipo de dimensão de mudança lenta para cada atributo de dimensão. Além disso, o design detalhado da tabela de fatos deve identificar cada relacionamento de chave estrangeira, dimensões degeneradas apropriadas e regras para cada fato para indicar se é aditivo, semi-aditivo ou não aditivo.

A planilha de design dimensional é o primeiro passo para criar o documento de mapeamento de origem para destino. A equipe de design físico irá aprofundar o mapeamento com nomes físicos de tabelas e colunas, tipos de dados e declarações de chaves.

  
**Acompanhe Questões do Modelo**

Qualquer questão, definição, regra de transformação e desafio de qualidade de dados descobertos durante o processo de design devem ser capturados em um log de acompanhamento de questões. Alguém deve ser designado para capturar e acompanhar as questões durante as sessões; o gerente de projeto, se estiver participando das sessões de design, muitas vezes lida com essa responsabilidade porque geralmente são adeptos de manter a lista atualizada e incentivar o progresso na resolução de questões pendentes. O facilitador deve reservar tempo adequado no final de cada sessão para revisar e validar novas entradas de questões e suas atribuições. Entre as sessões de design, a equipe de design está normalmente ocupada perfilando dados, buscando clarificações e acordo sobre definições comuns e se reunindo com especialistas do sistema de origem para resolver questões pendentes.

**Mantenha a Matriz de Barramento Atualizada**

Durante o processo de modelagem detalhada, frequentemente há novas descobertas sobre o processo de negócios que está sendo modelado. Frequentemente, essas descobertas resultam na introdução de novas tabelas de fatos para suportar o processo de negócios, novas dimensões ou a divisão ou combinação de dimensões. Você deve manter a matriz de barramento atualizada ao longo do processo de design porque é uma ferramenta chave de comunicação e planejamento. Como discutido no Capítulo 16: Seguros, a matriz de barramento detalhada frequentemente captura informações adicionais sobre a granularidade e métricas de cada tabela de fatos.

**Revise e Valide o Modelo**

Uma vez que a equipe de design esteja confiante sobre o modelo, o processo se move para a fase de revisão e validação para obter feedback de outras partes interessadas, incluindo:

* Recursos de TI, como membros da equipe DW/BI que não estão envolvidos no esforço de modelagem, especialistas no sistema de origem e DBAs
* Usuários analíticos ou avançados de negócios que não estão envolvidos no esforço de modelagem
* Comunidade mais ampla de usuários de negócios

**Revisões de TI**

Normalmente, a primeira revisão do modelo dimensional detalhado é com pares na organização de TI. Este público é frequentemente composto de revisores que estão intimamente familiarizados com o processo de negócios alvo porque escreveram ou gerenciam o sistema que o executa. Eles também estão pelo menos parcialmente familiarizados com o modelo de dados alvo porque você já tem estado incomodando-os com perguntas sobre dados de origem.

As revisões de TI podem ser desafiadoras porque os participantes frequentemente não têm um entendimento da modelagem dimensional. Na verdade, a maioria deles provavelmente se considera proficiente em modelagem 3NF. Sua tendência será aplicar regras de modelagem orientadas ao processamento de transações ao modelo dimensional. Em vez de gastar a maior parte do seu tempo debatendo os méritos de diferentes disciplinas de modelagem, é melhor fornecer proativamente alguma educação sobre modelagem dimensional como parte do processo de revisão.

Quando todos tiverem os conceitos básicos, você deve começar com uma revisão da matriz de barramento. Isso dá a todos uma sensação do escopo do projeto e da arquitetura geral de dados, demonstra o papel das dimensões conformadas e mostra as prioridades dos processos de negócios relativos. Em seguida, ilustre como a linha selecionada na matriz se traduz diretamente no diagrama de modelo dimensional de alto nível. Isso dá a todos o mapa em nível de entidade do modelo e serve como guia para o resto da discussão.

A maior parte da sessão de revisão deve ser gasta passando pelos detalhes das planilhas de dimensões e tabelas de fatos. Também é uma boa ideia revisar quaisquer questões abertas restantes para cada tabela enquanto você trabalha no modelo.

Mudanças no modelo provavelmente resultarão desta reunião. Lembre-se de atribuir a tarefa de capturar as questões e recomendações a alguém da equipe.

**Revisão dos Usuários Centrais**

Em muitos projetos, essa revisão não é necessária porque os principais usuários de negócios são membros da equipe de modelagem e já estão intimamente familiarizados com o modelo dimensional. Caso contrário, esta reunião de revisão é semelhante em escopo e estrutura à reunião de revisão de TI. Os principais usuários de negócios são mais técnicos do que os usuários de negócios típicos e podem lidar com detalhes sobre o modelo. Em organizações menores, frequentemente combinamos a revisão de TI e a revisão dos usuários centrais em uma única sessão.

**Revisão da Comunidade Mais Ampla de Usuários de Negócios**

Esta sessão é tanto uma educação quanto uma revisão de design. Você quer educar as pessoas sem sobrecarregá-las, enquanto ao mesmo tempo ilustra como o modelo dimensional suporta seus requisitos de negócios. Você deve começar com a matriz de barramento como o roteiro de dados DW/BI empresarial, revisar os diagramas de modelo de alto nível, e finalmente, revisar as dimensões críticas, como cliente e produto. Às vezes, os diagramas de bolhas são complementados com diagramas semelhantes ao da Figura 18-4 para ilustrar os caminhos de drill hierárquicos dentro de uma dimensão.

Certifique-se de alocar tempo durante esta educação/revisão para ilustrar como o modelo pode ser usado para responder a uma ampla gama de perguntas sobre o processo de negócios. Frequentemente, pegamos alguns exemplos do documento de requisitos e passamos por como eles seriam respondidos.

**Finalize a Documentação do Design**

Depois que o modelo estiver em sua forma final, a documentação do design deve ser compilada a partir dos documentos de trabalho da equipe de design. Este documento tipicamente inclui:

* Breve descrição do projeto
* Diagrama do modelo de dados de alto nível
* Planilha de design dimensional detalhada para cada tabela de fatos e dimensões
* Questões em aberto

**Resumo**  
  
 A modelagem dimensional é um processo de design iterativo que exige o esforço cooperativo de pessoas com um conjunto diversificado de habilidades, incluindo representantes de negócios. O esforço de design começa com um modelo gráfico inicial extraído da matriz de barramento e apresentado no nível de entidade. O processo de modelagem detalhada aprofunda-se nas definições, fontes, relacionamentos, problemas de qualidade de dados e transformações necessárias para cada tabela. Os objetivos principais são criar um modelo que atenda aos requisitos de negócios, verificar se os dados estão disponíveis para popular o modelo e fornecer à equipe de ETL uma direção clara.

A tarefa de determinar nomes de colunas e tabelas está entrelaçada no processo de design. A organização como um todo deve concordar com os nomes, definições e derivações de cada coluna e tabela no modelo dimensional. Este é mais um processo político do que técnico, que exige a atenção plena do membro mais diplomático da equipe. Os nomes de colunas resultantes expostos pela ferramenta de BI devem fazer sentido para a comunidade de negócios.

O esforço de modelagem detalhada é seguido por várias revisões. O resultado final é um modelo dimensional que foi testado com sucesso contra as necessidades de negócios e as realidades dos dados.

****Capítulo 19****

### ****Subsistemas e Técnicas de ETL****

O sistema de extração, transformação e carga (ETL) consome uma parte desproporcional do tempo e esforço necessários para construir um ambiente DW/BI. Desenvolver o sistema ETL é desafiador devido a muitas restrições externas que pressionam seu design: os requisitos de negócios, as realidades dos dados de origem, o orçamento, as janelas de processamento e as habilidades da equipe disponível. No entanto, pode ser difícil apreciar o porquê de o sistema ETL ser tão complexo e intensivo em recursos. Todos entendem as três letras: você extrai os dados de sua localização original (E), faz algo com eles (T) e, em seguida, carrega-os (L) em um conjunto final de tabelas para que os usuários de negócios possam consultar.

Quando perguntados sobre a melhor maneira de projetar e construir o sistema ETL, muitos designers respondem: “Bem, isso depende”. Depende da origem; depende das limitações dos dados; depende das linguagens de script e ferramentas ETL disponíveis; depende das habilidades da equipe; e depende das ferramentas de BI. Mas a resposta “depende” é perigosa porque se torna uma desculpa para adotar uma abordagem não estruturada ao desenvolver um sistema ETL, o que no pior cenário resulta em uma confusão de tabelas, módulos, processos, scripts, gatilhos, alertas e agendamentos de tarefas. Esta abordagem "criativa" de design não deve ser tolerada. Com a sabedoria retrospectiva de milhares de data warehouses bem-sucedidos, um conjunto de melhores práticas de ETL emergiu. Não há razão para tolerar uma abordagem não estruturada.

Uma consideração cuidadosa dessas melhores práticas revelou que 34 subsistemas são necessários em quase todo data warehouse dimensional. Não é de se admirar que o sistema ETL consuma uma porcentagem tão grande dos recursos de desenvolvimento de DW/BI!

Este capítulo é retirado de "The Data Warehouse Lifecycle Toolkit, Second Edition" (Wiley, 2008). Ao longo do capítulo, distribuímos referências para recursos no site do Kimball Group para uma cobertura mais aprofundada de várias técnicas de ETL. O Capítulo 19 revisa os seguintes conceitos:

* Requisitos e restrições a serem considerados antes de projetar o sistema ETL
* Três subsistemas focados em extrair dados de sistemas de origem
* Cinco subsistemas para lidar com limpeza e conformidade de valor agregado, incluindo estruturas dimensionais para monitorar erros de qualidade
* Treze subsistemas para entregar dados em estruturas dimensionais já conhecidas, como um subsistema para implementar técnicas de dimensões lentamente mutáveis
* Treze subsistemas para ajudar a gerenciar o ambiente de produção do ETL

### Coletar os Requisitos

Estabelecer a arquitetura de um sistema ETL começa com um dos desafios mais difíceis: reunir os requisitos. Com isso, queremos dizer coletar e entender todos os requisitos, realidades e restrições conhecidos que afetam o sistema ETL. A lista de requisitos pode ser bastante esmagadora, mas é essencial colocá-los na mesa antes de iniciar o desenvolvimento do sistema ETL. Os requisitos do sistema ETL são, em sua maioria, restrições com as quais você deve conviver e adaptar seu sistema. Dentro da estrutura desses requisitos, há oportunidades para tomar suas próprias decisões, exercer julgamento e alavancar a criatividade, mas os requisitos ditam os elementos essenciais que o sistema ETL deve entregar. As seguintes dez seções descrevem as principais áreas de requisitos que impactam o design e desenvolvimento do sistema ETL.

Antes de iniciar o esforço de design e desenvolvimento do ETL, você deve fornecer uma resposta curta para cada um dos seguintes dez requisitos. Fornecemos uma lista de verificação de amostra (como uma nota) para cada um para ajudá-lo a começar. O objetivo deste exercício é garantir que você visite cada um desses tópicos, pois qualquer um deles pode parar o projeto em algum momento.

### Necessidades de Negócio

Do ponto de vista de um designer de ETL, as necessidades de negócios são os requisitos de informação dos usuários do sistema DW/BI. Usamos o termo necessidades de negócios de forma um tanto restrita aqui para significar o conteúdo da informação que os usuários de negócios precisam para tomar decisões informadas. Como as necessidades de negócios direcionam diretamente a escolha das fontes de dados e sua subsequente transformação no sistema ETL, a equipe de ETL deve entender e examinar cuidadosamente as necessidades de negócios.

**NOTA:** Você deve manter uma lista dos principais indicadores de desempenho (KPIs) descobertos durante a definição dos requisitos de negócios que o projeto pretende suportar, bem como os objetivos de drill-down e drill-across necessários quando um usuário de negócios precisa investigar o motivo de um KPI ter mudado.

### Conformidade

Requisitos legais e de relatórios em constante mudança forçaram muitas organizações a apertar seriamente seus relatórios e fornecer prova de que os números relatados são precisos, completos e não foram adulterados. Claro, sistemas DW/BI em negócios regulamentados, como telecomunicações, cumprem requisitos de relatórios regulatórios há anos. Mas certamente o tom de relatórios financeiros se tornou muito mais rigoroso para todos.

**NOTA:** Em consulta com seu departamento jurídico ou diretor de conformidade (se você tiver um!) e a equipe de entrega de BI, você deve listar todos os dados e relatórios finais sujeitos a restrições de conformidade. Liste as entradas de dados e etapas de transformação de dados para as quais você deve manter a "cadeia de custódia" mostrando e provando que os relatórios finais foram derivados dos dados originais entregues de suas fontes de dados. Liste os dados para os quais você deve fornecer prova de segurança para as cópias sob seu controle, tanto offline quanto online. Liste as cópias de dados que você deve arquivar e a expectativa de vida útil desses arquivos. Boa sorte com tudo isso. É por isso que você é tão bem pago...

### Qualidade dos Dados

Três forças poderosas convergiram para colocar as preocupações com a qualidade dos dados no topo da lista para os executivos. Primeiro, a tendência cultural de longo prazo que diz: "Se ao menos eu pudesse ver os dados, então eu poderia gerenciar melhor o meu negócio" continua a crescer; os trabalhadores do conhecimento de hoje acreditam instintivamente que os dados são um requisito crucial para que eles funcionem em seus empregos. Segundo, a maioria das organizações entende que suas fontes de dados são profundamente distribuídas, tipicamente ao redor do mundo, e que integrar efetivamente uma miríade de fontes de dados díspares é necessário. E terceiro, as demandas aumentadas drasticamente por conformidade significam que o manuseio descuidado dos dados não será negligenciado ou desculpado.

**NOTA:** Você deve listar os elementos de dados cuja qualidade é conhecida por ser inaceitável e listar se um acordo foi alcançado com os sistemas de origem para corrigir os dados antes da extração. Liste os elementos de dados descobertos durante o perfilamento de dados, que serão continuamente monitorados e sinalizados como parte do processo ETL.

### Segurança

A conscientização sobre segurança aumentou significativamente nos últimos anos em toda a TI, mas muitas vezes continua sendo um pensamento tardio e um fardo indesejado para a maioria das equipes DW/BI. Os ritmos básicos do data warehouse estão em desacordo com a mentalidade de segurança; o data warehouse busca publicar dados amplamente para os tomadores de decisão, enquanto os interesses de segurança assumem que os dados devem ser restritos àqueles que precisam saber. Além disso, a segurança deve ser estendida aos backups físicos. Se a mídia puder ser facilmente removida do cofre de backup, então a segurança foi comprometida tão efetivamente quanto se as senhas online fossem comprometidas.

Durante a coleta de requisitos, a equipe DW/BI deve buscar orientação clara da alta administração sobre quais aspectos do sistema DW/BI carregam sensibilidade extra de segurança. Se essas questões nunca foram examinadas, é provável que a questão seja devolvida à equipe. Esse é o momento em que um gerente de segurança experiente deve ser convidado a se juntar à equipe de design. Os requisitos de conformidade provavelmente se sobrepõem aos requisitos de segurança; pode ser sábio combinar esses dois tópicos durante a coleta de requisitos.

**NOTA:** Você deve expandir a lista de verificação de conformidade para englobar requisitos conhecidos de segurança e privacidade.

### Integração de Dados

A integração de dados é um grande tópico para a TI porque, em última análise, visa fazer com que todos os sistemas funcionem perfeitamente juntos. A "visão de 360 graus da empresa" é um nome familiar para a integração de dados. Em muitos casos, a integração séria de dados deve ocorrer entre os sistemas de transações primários da organização antes que os dados cheguem à porta dos fundos do data warehouse. Mas raramente essa integração de dados é completa, a menos que a organização tenha um sistema abrangente e centralizado de gerenciamento de dados mestres (MDM), e mesmo assim é provável que outros sistemas operacionais importantes existam fora do sistema MDM primário.

A integração de dados geralmente assume a forma de conformação de dimensões e fatos no data warehouse. Conformar dimensões significa estabelecer atributos dimensionais comuns entre bancos de dados separados, para que relatórios de drill-across possam ser gerados usando esses atributos. Conformar fatos significa fazer acordos sobre métricas de negócios comuns, como indicadores-chave de desempenho (KPIs) entre bancos de dados separados, para que esses números possam ser comparados matematicamente calculando diferenças e razões.

**NOTA:** Você deve usar a matriz de negócios para gerar uma lista de prioridade para a conformação de dimensões (colunas da matriz de negócios). Anote cada linha da matriz de negócios com se há uma demanda executiva clara para que o processo de negócios participe do processo de integração e se a equipe ETL responsável por aquele processo de negócios concordou.

### Latência de Dados

Latência de dados descreve quão rapidamente os dados do sistema de origem devem ser entregues aos usuários de negócios via sistema DW/BI. Obviamente, os requisitos de latência de dados têm um grande impacto na arquitetura do ETL. Algoritmos de processamento inteligentes, paralelização e hardware potente podem acelerar fluxos de dados orientados por lotes tradicionais. Mas em algum momento, se o requisito de latência de dados for suficientemente urgente, a arquitetura do sistema ETL deve converter de lote para micro-lote ou orientada por streaming. Esta mudança não é gradual ou evolutiva; é uma mudança de paradigma em que quase todos os passos do pipeline de entrega de dados devem ser reimplementados.

**NOTA:** Você deve listar todas as demandas de negócios legítimas e bem avaliadas para dados que devem ser fornecidos diariamente, várias vezes ao dia, dentro de alguns segundos ou instantaneamente. Anote cada demanda com se a comunidade de negócios entende as concessões de qualidade de dados associadas à sua escolha particular. Perto do final do Capítulo 20: Processo e Tarefas de Design e Desenvolvimento do Sistema ETL, discutimos as concessões de qualidade de dados causadas por requisitos de baixa latência.

### Arquivamento e Linhagem

Requisitos de arquivamento e linhagem foram sugeridos nas seções anteriores de conformidade e segurança. Mesmo sem os requisitos legais para salvar dados, todo data warehouse precisa de várias cópias de dados antigos, seja para comparações com novos dados para gerar registros de captura de mudanças ou reprocessamento. Recomendamos estagiar os dados (escrevê-los em disco) após cada atividade importante do pipeline ETL: depois de serem extraídos, limpos e conformados, e entregues.

Então, quando o estagiamento se transforma em arquivamento onde os dados são mantidos indefinidamente em alguma forma de mídia permanente? Nossa resposta simples é uma resposta conservadora. Todos os dados estagiados devem ser arquivados, a menos que uma decisão consciente seja tomada de que conjuntos de dados específicos nunca serão recuperados no futuro. Quase sempre é menos problemático ler os dados da mídia permanente do que reprocessar os dados pelo sistema ETL em um momento posterior. E, claro, pode ser impossível reprocessar os dados de acordo com os antigos algoritmos de processamento se tempo suficiente tiver passado ou se a extração original não puder ser recriada.

E, enquanto estamos nisso, cada conjunto de dados estagiado/arquivado deve ter metadados acompanhando descrevendo as origens e etapas de processamento que produziram os dados. Novamente, o rastreamento desta linhagem é explicitamente exigido por certos requisitos de conformidade, mas deve fazer parte de todas as situações de arquivamento.

**NOTA:** Você deve listar as fontes de dados e etapas de dados intermediárias que serão arquivadas, juntamente com as políticas de retenção e restrições de conformidade, segurança e privacidade.

### Interfaces de Entrega de BI

A etapa final para o sistema ETL é a transferência para os aplicativos de BI. Adotamos uma posição forte e disciplinada sobre essa transferência. Acreditamos que a equipe ETL, trabalhando em estreita colaboração com a equipe de modelagem, deve assumir a responsabilidade pelo conteúdo e estrutura dos dados que tornam os aplicativos de BI simples e rápidos. Esta atitude é mais do que uma declaração vago. Acreditamos que é irresponsável transferir dados para o aplicativo de BI de uma maneira que aumente a complexidade do aplicativo, desacelere a criação de consultas ou relatórios ou torne os dados desnecessariamente complexos para os usuários de negócios. O erro mais elementar e sério é transferir um modelo físico completamente normalizado e abandonar o trabalho. É por isso que fazemos tanto esforço para construir estruturas dimensionais que compreendem a transferência final.

A equipe ETL e os modeladores de dados precisam trabalhar de perto com os desenvolvedores de aplicativos de BI para determinar os requisitos exatos para a transferência de dados. Cada ferramenta de BI tem certas sensibilidades que devem ser evitadas e certos recursos que podem ser explorados se os dados físicos estiverem no formato certo. As mesmas considerações se aplicam aos dados preparados para cubos OLAP.

**NOTA:** Você deve listar todas as tabelas de fato e dimensão que serão diretamente expostas às suas ferramentas de BI. Isso deve vir diretamente da especificação do modelo dimensional. Liste todos os cubos OLAP e estruturas de banco de dados especiais necessárias pelas ferramentas de BI. Liste todos os índices conhecidos e agregações que você concordou em construir para suportar o desempenho de BI.

### Habilidades Disponíveis

Algumas decisões de design do sistema ETL devem ser tomadas com base nos recursos disponíveis para construir e gerenciar o sistema. Você não deve construir um sistema que dependa de módulos de processamento C++ críticos se essas habilidades de programação não estiverem na empresa ou não puderem ser adquiridas razoavelmente. Da mesma forma, você pode estar muito mais confiante em construir o sistema ETL em torno de uma ferramenta ETL de um grande fornecedor se você já tiver essas habilidades internamente e souber como gerenciar tal projeto.

Considere a grande decisão de codificar manualmente o sistema ETL ou usar um pacote ETL de um fornecedor. Questões técnicas e custos de licença à parte, não vá em uma direção que seus funcionários e gerentes considerem desconhecida sem considerar seriamente as implicações de longo prazo da decisão.

**NOTA:** Você deve inventariar o sistema operacional do seu departamento, ferramenta ETL, linguagem de script, linguagem de programação, habilidades em SQL, DBMS e OLAP para entender quão exposto você está a uma escassez ou perda dessas habilidades. Liste essas habilidades necessárias para suportar seus sistemas atuais e seus sistemas futuros prováveis.

### Licenças Legadas

Finalmente, em muitos casos, grandes decisões de design serão feitas implicitamente pela insistência da alta administração em usar licenças legadas existentes. Em muitos casos, esse requisito é algo com o qual você pode conviver porque as vantagens ambientais são claras para todos. Mas em alguns casos, o uso de uma licença legada para o desenvolvimento ETL é um erro. Esta é uma posição difícil de estar e, se você sentir fortemente o suficiente, pode precisar apostar seu emprego. Se precisar abordar a alta administração e desafiar o uso de uma licença legada existente, esteja bem preparado para defender o caso e esteja disposto a aceitar a decisão final ou possivelmente procurar emprego em outro lugar.

**NOTA:** Você deve listar seu sistema operacional legado, ferramenta ETL, linguagem de script, linguagem de programação, SQL, DBMS e licenças OLAP e se seu uso exclusivo é obrigatório ou apenas recomendado.

**Os 34 Subsistemas de ETL**

Com um entendimento dos requisitos, realidades e restrições existentes, você está pronto para aprender sobre os 34 subsistemas críticos que formam a arquitetura de todo sistema ETL. Este capítulo descreve todos os 34 subsistemas com igual ênfase. O próximo capítulo descreve os passos práticos para implementar esses subsistemas necessários para cada situação específica. Embora tenhamos adotado o vernáculo da indústria, ETL, para descrever esses passos, o processo realmente possui quatro componentes principais:

* **Extração**: Coletando dados brutos dos sistemas de origem e geralmente gravando-os em disco no ambiente ETL antes de qualquer reestruturação significativa dos dados ocorrer. Os subsistemas de 1 a 3 apoiam o processo de extração.
* **Limpeza e Conformidade**: Enviando dados de origem através de uma série de etapas de processamento no sistema ETL para melhorar a qualidade dos dados recebidos da fonte, e mesclando dados de duas ou mais fontes para criar e impor dimensões conformadas e métricas conformadas. Os subsistemas de 4 a 8 descrevem a arquitetura necessária para apoiar os processos de limpeza e conformidade.
* **Entrega**: Estruturando fisicamente e carregando os dados nos modelos dimensionais de destino do servidor de apresentação. Os subsistemas de 9 a 21 fornecem as capacidades para entregar os dados ao servidor de apresentação.
* **Gerenciamento**: Gerenciando os sistemas e processos relacionados do ambiente ETL de maneira coerente. Os subsistemas de 22 a 34 descrevem os componentes necessários para apoiar o gerenciamento contínuo do sistema ETL.

**Extração: Obtendo Dados no Data Warehouse**

Não é surpresa que os subsistemas iniciais da arquitetura ETL abordem as questões de entender seus dados de origem, extrair os dados e transferi-los para o ambiente do data warehouse onde o sistema ETL pode operar de forma independente dos sistemas operacionais. Embora os subsistemas restantes se concentrem na transformação, no carregamento e na gestão do sistema dentro do ambiente ETL, os subsistemas iniciais fazem interface com os sistemas de origem para acesso aos dados necessários.

**Subsistema 1: Perfilamento de Dados**

O perfilamento de dados é a análise técnica dos dados para descrever seu conteúdo, consistência e estrutura. Em certo sentido, sempre que você realiza uma consulta investigativa SELECT DISTINCT em um campo de banco de dados, você está fazendo perfilamento de dados. Há uma variedade de ferramentas especificamente projetadas para fazer um perfilamento poderoso. Provavelmente vale a pena investir em uma ferramenta em vez de desenvolver a sua própria porque as ferramentas permitem que muitas relações de dados sejam facilmente exploradas com simples gestos da interface do usuário. Você pode ser muito mais produtivo nas etapas de perfilamento de dados de um projeto usando uma ferramenta em vez de codificar manualmente todas as perguntas sobre o conteúdo dos dados.

O perfilamento de dados desempenha dois papéis distintos: estratégico e tático. Assim que uma fonte de dados candidata é identificada, deve ser feita uma avaliação de perfilamento leve para determinar sua adequação para inclusão no data warehouse e fornecer uma decisão inicial de prosseguir ou não. Idealmente, essa avaliação estratégica deve ocorrer imediatamente após a identificação de uma fonte de dados candidata durante a análise de requisitos de negócios. A desqualificação precoce de uma fonte de dados é um passo responsável que pode lhe render respeito do resto da equipe, mesmo que seja uma má notícia. Uma revelação tardia de que a fonte de dados não suporta a missão pode descarrilar a iniciativa DW/BI (e ser um desfecho de carreira potencialmente fatal para você), especialmente se essa revelação ocorrer meses após o início de um projeto.

Após a decisão estratégica básica de incluir uma fonte de dados no projeto, um esforço tático de perfilamento de dados deve ocorrer para eliminar tantos problemas quanto possível. Geralmente, essa tarefa começa durante o processo de modelagem de dados e se estende para o processo de design do sistema ETL. Às vezes, espera-se que a equipe ETL inclua uma fonte com conteúdo que não foi completamente avaliado. Os sistemas podem suportar as necessidades dos processos de produção, mas apresentam desafios para o ETL, porque campos que não são centrais para o processamento de produção podem ser pouco confiáveis e incompletos para fins de análise. Problemas que surgem neste subsistema resultam em especificações detalhadas que são enviadas de volta ao originador da fonte de dados como solicitações de melhoria ou formam requisitos para o processamento de qualidade de dados descrito nos subsistemas de 4 a 8.

A etapa de perfilamento fornece à equipe ETL orientações sobre quanto maquinário de limpeza de dados deve ser invocado e protege-os de perder grandes marcos do projeto devido ao desvio inesperado de construir sistemas para lidar com dados sujos. Faça o perfilamento de dados com antecedência! Use os resultados do perfilamento de dados para definir as expectativas dos patrocinadores do negócio sobre cronogramas de desenvolvimento realistas, limitações nos dados de origem e a necessidade de investir em práticas melhores de captura de dados de origem.

**Subsistema 2: Sistema de Captura de Mudanças de Dados**

Durante a carga histórica inicial do data warehouse, capturar mudanças no conteúdo dos dados de origem não é importante porque você carrega todos os dados a partir de um ponto no tempo para frente. No entanto, muitas tabelas do data warehouse são tão grandes que não podem ser atualizadas durante cada ciclo ETL. Você deve ter uma capacidade de transferir apenas as mudanças relevantes nos dados de origem desde a última atualização. Isolar os dados de origem mais recentes é chamado de captura de mudanças de dados (CDC). A ideia por trás do CDC é simples o suficiente: apenas transferir os dados que mudaram desde a última carga. Mas construir um bom sistema de CDC não é tão fácil quanto parece. Os principais objetivos para o subsistema de captura de mudanças de dados são:

* Isolar os dados de origem alterados para permitir processamento seletivo em vez de uma atualização completa.
* Capturar todas as mudanças (exclusões, edições e inserções) feitas nos dados de origem, incluindo mudanças feitas através de interfaces não padronizadas.
* Marcar dados alterados com códigos de motivo para distinguir correções de erros de atualizações verdadeiras.
* Apoiar o rastreamento de conformidade com metadados adicionais.
* Realizar a etapa de CDC o mais cedo possível, de preferência antes de uma transferência em massa de dados para o data warehouse.

Capturar mudanças de dados está longe de ser uma tarefa trivial. Você deve avaliar cuidadosamente sua estratégia para cada fonte de dados. Determinar a estratégia apropriada para identificar dados alterados pode exigir algum trabalho de detetive. As tarefas de perfilamento de dados descritas anteriormente podem ajudar a equipe ETL a fazer essa determinação. Há várias maneiras de capturar mudanças nos dados de origem, cada uma eficaz na situação apropriada, incluindo:

* Colunas de Auditoria Em alguns casos, o sistema de origem inclui colunas de auditoria que armazenam a data e a hora em que um registro foi adicionado ou modificado. Essas colunas geralmente são preenchidas por meio de gatilhos de banco de dados que são disparados automaticamente conforme os registros são inseridos ou atualizados. Às vezes, por razões de desempenho, as colunas são preenchidas pelo aplicativo de origem em vez de gatilhos de banco de dados. Quando esses campos são carregados por qualquer meio que não seja gatilhos de banco de dados, preste atenção especial à sua integridade, analisando e testando cada coluna para garantir que seja uma fonte confiável para indicar mudança. Se você descobrir quaisquer valores NULOS, você deve encontrar uma abordagem alternativa para detectar mudanças. A situação mais comum que impede o sistema ETL de usar colunas de auditoria é quando os campos são preenchidos pelo aplicativo de origem, mas a equipe de DBA permite que scripts de back-end modifiquem dados. Se isso ocorrer em seu ambiente, você enfrenta um alto risco de perder dados alterados durante as cargas incrementais. Finalmente, você precisa entender o que acontece quando um registro é excluído da fonte porque consultar a coluna de auditoria pode não capturar esse evento.
* Extrações Programadas Com uma extração programada, você normalmente seleciona todas as linhas onde os campos de data de criação ou modificação são iguais a SYSDATE-1, ou seja, todos os registros de ontem. Parece perfeito, certo? Errado. Carregar registros baseados puramente no tempo é um erro comum cometido por desenvolvedores ETL inexperientes. Este processo é terrivelmente confiável. A seleção de dados baseada no tempo carrega linhas duplicadas quando é reiniciada a partir de falhas no processo intermediário. Isso significa que a intervenção manual e a limpeza de dados são necessárias se o processo falhar por qualquer motivo. Enquanto isso, se o processo de carga noturna falhar em ser executado e pular um dia, há um risco de que os dados perdidos nunca cheguem ao data warehouse.
* Comparação de Diferença Completa Uma comparação de diferença completa mantém um instantâneo completo dos dados de ontem e os compara, registro por registro, com os dados de hoje para encontrar o que mudou. A boa notícia é que essa técnica é completa: você está garantido a encontrar toda mudança. A má notícia óbvia é que, em muitos casos, essa técnica é muito intensiva em recursos. Se uma comparação de diferença completa for necessária, tente fazer a comparação na máquina de origem, para que você não tenha que transferir a tabela inteira ou o banco de dados para o ambiente ETL. Claro, as pessoas de suporte de origem podem ter uma opinião sobre isso. Além disso, investigue o uso de algoritmos de soma de verificação de redundância cíclica (CRC) para dizer rapidamente se um registro complexo mudou sem examinar cada campo individualmente.
* Raspagem de Log de Banco de Dados A raspagem de log efetivamente tira uma foto do log de reexecução do banco de dados em um ponto agendado no tempo (geralmente à meia-noite) e o vasculha em busca de transações que afetam as tabelas de interesse para a carga ETL. A checagem envolve uma sondagem do log de reexecução, capturando transações em tempo real. Raspagem do log para transações é provavelmente a técnica mais bagunçada de todas. Não é incomum que os logs de transações fiquem cheios e impeçam o processamento de novas transações. Quando isso acontece em um ambiente de transação de produção, a reação instintiva do DBA responsável pode ser esvaziar o log para que as operações de negócios possam ser retomadas, mas quando um log é esvaziado, todas as transações dentro dele são perdidas. Se você esgotou todas as outras técnicas e descobriu que a raspagem de log é seu último recurso para encontrar registros novos ou alterados, convença o DBA a criar um log especial para atender às suas necessidades específicas.
* Monitoramento de Fila de Mensagens Em um sistema de transação baseado em mensagens, a fila é monitorada para todas as transações contra as tabelas de interesse. O conteúdo da transmissão é semelhante ao que você obtém com a checagem de log. Um benefício desse processo é a sobrecarga relativamente baixa, assumindo que a fila de mensagens já esteja em vigor. No entanto, pode não haver um recurso de replay na fila de mensagens. Se a conexão com a fila de mensagens for perdida, você perde dados.

**Subsistema 3: Sistema de Extração**

Obviamente, extrair dados dos sistemas de origem é um componente fundamental da arquitetura ETL. Se você tiver muita sorte, todos os dados de origem estarão em um único sistema que pode ser facilmente extraído usando uma ferramenta ETL. Na situação mais comum, cada fonte pode estar em um sistema, ambiente e/ou DBMS diferente. O sistema ETL pode ser esperado para extrair dados de uma ampla variedade de sistemas envolvendo muitos tipos diferentes de dados e desafios inerentes. Organizações que precisam extrair dados de ambientes de mainframe frequentemente encontram problemas envolvendo livros de cópias COBOL, conversões de EBCDIC para ASCII, decimais compactados, redefinições, campos OCCURS e tipos de registros múltiplos e variáveis. Outras organizações podem precisar extrair de fontes em DBMS relacionais, arquivos planos, fontes XML, registros da web ou um sistema ERP complexo. Cada um apresenta uma variedade de desafios possíveis. Algumas fontes, especialmente sistemas legados mais antigos, podem exigir o uso de linguagens procedurais diferentes das que a ferramenta ETL pode suportar ou com as quais a equipe tem experiência. Nessa situação, solicite que o proprietário do sistema de origem extraia os dados em um formato de arquivo plano.

**NOTA** Embora os dados formatados em XML tenham muitas vantagens porque são autoexplicativos, você pode não querê-los para transferências de dados grandes e frequentes. A porção de carga útil de um arquivo típico formatado em XML pode ser menos de 10 porcento do arquivo total. A exceção a essa recomendação pode ser onde a carga útil XML é uma estrutura XML hierárquica profundamente complexa, como uma troca de dados padrão da indústria. Nesses casos, a equipe DW/BI deve decidir se deve "desmembrar" o XML em um grande número de tabelas de destino ou persistir a estrutura XML dentro do data warehouse. Avanços recentes no suporte de fornecedores de RDBMS para XML via XPath tornaram essa última opção viável.

Há dois métodos primários para obter dados de um sistema de origem: como um arquivo ou um fluxo. Se a fonte é um sistema de mainframe envelhecido, geralmente é mais fácil extrair para arquivos e depois mover esses arquivos para o servidor ETL.

**NOTA** Se os dados de origem são não estruturados, semi-estruturados ou até mesmo "big data" hiperestruturados, então, em vez de carregar esses dados como um "blob" RDBMS ininterpretável, muitas vezes é mais eficaz criar uma etapa de extração MapReduce/Hadoop que funcione como um extrator de fatos ETL dos dados de origem, entregando diretamente dados RDBMS carregáveis.

Se você usar uma ferramenta ETL e os dados de origem estiverem em um banco de dados (não necessariamente um RDBMS), você pode configurar a extração como um fluxo onde os dados fluem para fora do sistema de origem, através do motor de transformação, e para o banco de dados de staging como um único processo. Em contraste, uma abordagem de extração para arquivo consiste em três ou quatro etapas discretas: extrair para o arquivo, mover o arquivo para o servidor ETL, transformar o conteúdo do arquivo e carregar os dados transformados no banco de dados de staging.

**NOTA** Embora a extração de fluxo seja mais atraente, as extrações para arquivo têm algumas vantagens. Eles são fáceis de reiniciar em vários pontos. Desde que você salve o arquivo de extração, você pode rerodar a carga sem impactar o sistema de origem. Você pode facilmente criptografar e comprimir os dados antes de transferi-los pela rede. Finalmente, é fácil verificar se todos os dados foram movidos corretamente comparando as contagens de linhas do arquivo antes e depois da transferência. Geralmente, recomendamos uma utilidade de transferência de dados, como FTP, para mover o arquivo extraído.

A compressão de dados é importante se grandes quantidades de dados precisam ser transferidas por uma distância significativa ou através de uma rede pública. Nesse caso, o link de comunicação é frequentemente o gargalo. Se muito tempo for gasto na transmissão dos dados, a compressão pode reduzir o tempo de transmissão em 30 a 50 porcento ou mais, dependendo da natureza do arquivo de dados original.

A criptografia de dados é importante se os dados forem transferidos através de uma rede pública, ou mesmo internamente em algumas situações. Se esse for o caso, é melhor enviar tudo através de um link criptografado e não se preocupar com o que precisa ser seguro e o que não precisa. Lembre-se de comprimir antes de criptografar porque arquivos criptografados não comprimem muito bem.

**Limpeza e Conformidade de Dados**

A limpeza e conformidade de dados são tarefas críticas do sistema ETL. Essas são as etapas onde o sistema ETL adiciona valor aos dados. As outras atividades, extração e entrega de dados, são obviamente necessárias, mas simplesmente movem e carregam os dados. Os subsistemas de limpeza e conformidade realmente alteram os dados e aumentam seu valor para a organização. Além disso, esses subsistemas podem ser arquitetados para criar metadados usados para diagnosticar o que está errado com os sistemas de origem. Tais diagnósticos podem eventualmente levar a iniciativas de reengenharia de processos de negócios para abordar as causas raízes de dados sujos e melhorar a qualidade dos dados ao longo do tempo.

**Melhorando a Cultura e os Processos de Qualidade de Dados**

É tentador culpar a fonte de dados original por quaisquer e todos os erros que aparecem a jusante. Se apenas os digitadores de dados fossem mais cuidadosos! Somos apenas um pouco mais tolerantes com vendedores desafiados pelo teclado que inserem informações de clientes e produtos em seus formulários de pedido. Talvez você possa resolver o problema de qualidade de dados impondo restrições nas interfaces de entrada de dados do usuário. Esta abordagem fornece uma dica sobre como pensar em corrigir a qualidade dos dados porque uma solução técnica muitas vezes evita o problema real. Suponha que os campos do número de Seguro Social para clientes frequentemente estejam em branco ou preenchidos com lixo em uma tela de entrada. Alguém tem a brilhante ideia de exigir a entrada no formato 999-99-9999 e de proibir inteligentemente entradas sem sentido, como todos os 9s. O que acontece? Os digitadores de dados são forçados a fornecer números de Seguro Social válidos para avançar para a próxima tela, então, quando eles não têm o número do cliente, eles digitam um número artificial que passa pelo bloqueio.

Michael Hammer, em seu livro revolucionário Reengenharia da Corporação (Collins, revisado em 2003), atingiu o coração do problema de qualidade de dados com uma observação brilhante. Parafraseando Hammer: “Problemas de qualidade de dados aparentemente pequenos são, na realidade, indicações importantes de processos de negócios quebrados.” Não apenas essa percepção foca corretamente sua atenção na fonte dos problemas de qualidade de dados, mas também mostra o caminho para a solução.

Tentativas técnicas de abordar a qualidade dos dados não prevalecerão a menos que façam parte de uma cultura de qualidade geral que deve vir do topo de uma organização. A famosa atitude de qualidade na fabricação de carros japoneses permeia todos os níveis dessas organizações, e a qualidade é abraçada entusiasticamente por todos os níveis, do CEO ao trabalhador da linha de montagem. Para colocar isso em um contexto de dados, imagine uma empresa como uma grande rede de drogarias, onde uma equipe de compradores contrata com milhares de fornecedores para fornecer o inventário. Os compradores têm assistentes, cuja tarefa é inserir as descrições detalhadas de tudo o que é comprado pelos compradores. Essas descrições contêm dezenas de atributos. Mas o problema é que os assistentes têm um trabalho mortal e são julgados por quantos itens eles inserem por hora. Os assistentes têm quase nenhuma consciência de quem usa seus dados. Ocasionalmente, os assistentes são repreendidos por erros óbvios. Mas mais insidiosamente, os dados dados aos assistentes são eles mesmos incompletos e não confiáveis. Por exemplo, não há padrões formais para classificações de toxicidade, então há variação significativa ao longo do tempo e entre categorias de produtos para esse atributo. Como a drogaria melhora a qualidade dos dados? Aqui está um modelo de nove etapas, não apenas para a drogaria, mas para qualquer organização que aborde a qualidade dos dados:

* Declare um compromisso de alto nível com uma cultura de qualidade de dados.
* Conduza a reengenharia de processos no nível executivo.
* Gaste dinheiro para melhorar o ambiente de entrada de dados.
* Gaste dinheiro para melhorar a integração de aplicativos.
* Gaste dinheiro para mudar como os processos funcionam.
* Promova a conscientização de equipe de ponta a ponta.
* Promova a cooperação interdepartamental.
* Celebre publicamente a excelência em qualidade de dados.
* Meça e melhore continuamente a qualidade dos dados.

Na drogaria, dinheiro precisa ser gasto para melhorar o sistema de entrada de dados, para que ele forneça o conteúdo e as escolhas necessárias pelos assistentes dos compradores. Os executivos da empresa precisam garantir aos assistentes dos compradores que seu trabalho é importante e afeta muitos tomadores de decisão de maneira positiva. Os esforços diligentes dos assistentes devem ser publicamente elogiados e recompensados. E a conscientização e apreciação de equipe de ponta a ponta do valor comercial derivado de dados de qualidade é o objetivo final.

**Subsistema 4: Sistema de Limpeza de Dados**

O processo de limpeza de dados ETL é frequentemente esperado para corrigir dados sujos, mas ao mesmo tempo espera-se que o data warehouse forneça uma imagem precisa dos dados conforme foram capturados pelos sistemas de produção da organização. Encontrar o equilíbrio adequado entre esses objetivos conflitantes é essencial. Um de nossos objetivos ao descrever o sistema de limpeza é oferecer uma arquitetura abrangente para limpar dados, capturar eventos de qualidade de dados, bem como medir e, em última análise, controlar a qualidade dos dados no data warehouse. Algumas organizações podem achar essa arquitetura desafiadora de implementar, mas estamos convencidos de que é importante para a equipe ETL fazer um esforço sério para incorporar o máximo possível dessas capacidades. Se você é novo no ETL e acha isso um desafio assustador, você pode se perguntar: “Qual é o mínimo que devo focar?” A resposta é começar realizando a melhor análise de perfilamento de dados possível. Os resultados desse esforço podem ajudá-lo a entender os riscos de avançar com dados potencialmente sujos ou não confiáveis e ajudá-lo a determinar quão sofisticado seu sistema de limpeza de dados precisa ser.

O propósito dos subsistemas de limpeza é mobilizar a tecnologia para apoiar a qualidade dos dados. Os objetivos do subsistema devem incluir:

* Diagnóstico e triagem precoce de questões de qualidade de dados
* Requisitos para sistemas de origem e esforços de integração para fornecer dados melhores
* Fornecer descrições específicas de erros de dados esperados a serem encontrados no ETL
* Estrutura para capturar todos os erros de qualidade de dados e medir precisamente as métricas de qualidade de dados ao longo do tempo
* Anexar métricas de confiança de qualidade aos dados finais

**Telas de Qualidade**

O coração da arquitetura ETL é um conjunto de telas de qualidade que atuam como filtros diagnósticos nas tubulações de fluxo de dados. Cada tela de qualidade é um teste. Se o teste contra os dados for bem-sucedido, nada acontece e a tela não tem efeitos colaterais. Mas se o teste falhar, então deve descartar uma linha de evento de erro no esquema de evento de erro e escolher interromper o processo, enviar os dados ofensivos para suspensão ou simplesmente marcar os dados.

Embora todas as telas de qualidade sejam arquitetonicamente semelhantes, é conveniente dividi-las em três tipos, em ordem ascendente de escopo. Jack Olson, em seu livro seminal Qualidade de Dados: A Dimensão da Precisão (Morgan Kaufmann, 2002), classificou as telas de qualidade em três categorias: telas de coluna, telas de estrutura e telas de regra de negócios.

Telas de coluna testam os dados dentro de uma única coluna. Estes são geralmente testes simples, um tanto óbvios, como testar se uma coluna contém valores nulos inesperados, se um valor está fora de um intervalo prescrito ou se um valor falha em aderir a um formato exigido.

Telas de estrutura testam a relação dos dados entre colunas. Dois ou mais atributos podem ser testados para verificar se implementam uma hierarquia, como uma série de relacionamentos muitos-para-um. As telas de estrutura também testam relacionamentos de chave estrangeira/chave primária entre colunas em duas tabelas e também incluem testar blocos inteiros de colunas para verificar se implementam endereços postais válidos.

Telas de regra de negócios implementam testes mais complexos que não se encaixam nas categorias de tela de coluna ou estrutura mais simples. Por exemplo, um perfil de cliente pode ser testado por uma regra de negócios complexa dependente do tempo, como exigir que um passageiro frequente de platina vitalício tenha sido membro por pelo menos cinco anos e tenha voado mais de 2 milhões de milhas. As telas de regra de negócios também incluem verificações de qualidade de dados de limite agregado, como verificar se um número estatisticamente improvável de exames de ressonância magnética foi solicitado para diagnósticos menores como um cotovelo torcido. Nesse caso, a tela só lança um erro depois que um limite desses exames de ressonância magnética é alcançado.

**Respondendo a Eventos de Qualidade**

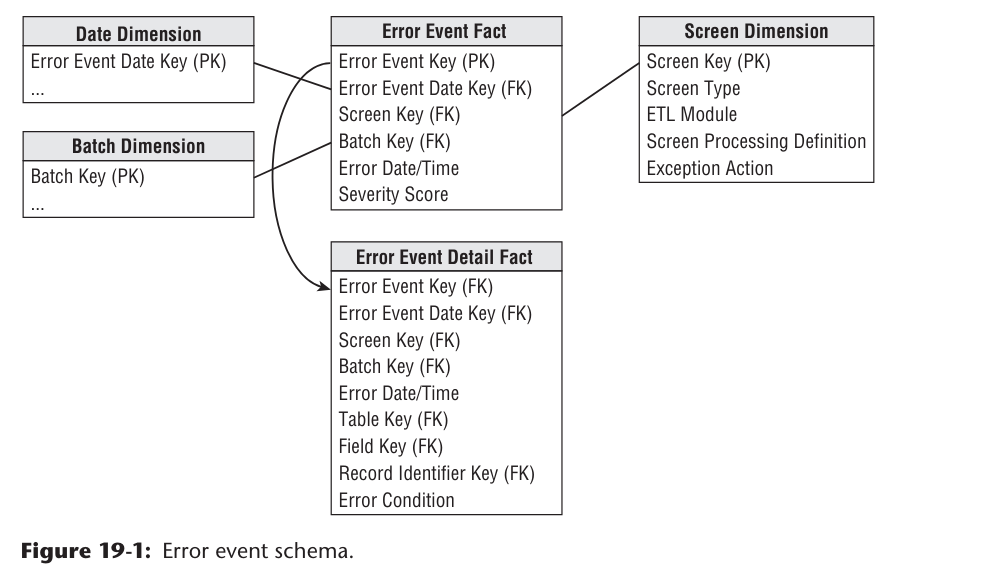
Já observamos que cada tela de qualidade tem que decidir o que acontece quando um erro é lançado. As escolhas são: 1) interromper o processo; 2) enviar o(s) registro(s) ofensivo(s) para um arquivo de suspense para processamento posterior; e 3) simplesmente marcar os dados e passá-los para a próxima etapa no pipeline. A terceira opção é de longe a melhor escolha, sempre que possível. Interromper o processo é obviamente uma dor porque requer intervenção manual para diagnosticar o problema, reiniciar ou retomar o trabalho ou abortar completamente. Enviar registros para um arquivo de suspense muitas vezes é uma solução ruim porque não está claro quando ou se esses registros serão corrigidos e reintroduzidos ao pipeline. Até que os registros sejam restaurados ao fluxo de dados, a integridade geral do banco de dados é questionável porque os registros estão faltando. Recomendamos não usar o arquivo de suspense para transgressões menores de dados. A terceira opção de marcar os dados com a condição de erro geralmente funciona bem. Dados ruins da tabela de fatos podem ser marcados com a dimensão de auditoria, conforme descrito no subsistema 6. Dados ruins da dimensão também podem ser marcados usando uma dimensão de auditoria ou, no caso de dados faltantes ou lixo, podem ser marcados com valores de erro únicos no próprio atributo.

**Subsistema 5: Esquema de Evento de Erro**

O esquema de evento de erro é um esquema dimensional centralizado cujo propósito é registrar cada evento de erro lançado por uma tela de qualidade em qualquer lugar na pipeline ETL. Embora nos concentremos no processamento ETL de data warehouse, essa abordagem pode ser usada em aplicações genéricas de integração de dados (DI) onde os dados estão sendo transferidos entre aplicações legadas. O esquema de evento de erro é mostrado na Figura 19-1.

A tabela principal é a tabela de fatos de evento de erro. Seu grão é cada erro lançado (produzido) por uma tela de qualidade em qualquer lugar no sistema ETL. Lembre-se de que o grão de uma tabela de fatos é a descrição física de por que uma linha da tabela de fatos existe. Assim, cada erro de tela de qualidade produz exatamente uma linha nesta tabela, e cada linha na tabela corresponde a um erro observado.

As dimensões da tabela de fatos de evento de erro incluem a data do calendário do erro, o trabalho em lote no qual o erro ocorreu e a tela que produziu o erro. A data do calendário não é um carimbo de data e hora de minuto e segundo do erro, mas sim fornece uma maneira de restringir e resumir eventos de erro pelos atributos usuais do calendário, como dia da semana ou último dia de um período fiscal. O fato data/hora do erro é um carimbo de data/hora relacional completo que especifica precisamente quando o erro ocorreu. Este formato é útil para calcular o intervalo de tempo entre eventos de erro porque você pode tomar a diferença entre dois carimbos de data/hora para obter o número de segundos que separam os eventos.



A dimensão do lote pode ser generalizada para ser uma etapa de processamento em casos em que os dados são transmitidos, em vez de processados em lotes. A dimensão da tela identifica precisamente qual é o critério da tela e onde reside o código para a tela. Ele também define o que fazer quando a tela lança um erro. (Por exemplo, interromper o processo, enviar o registro para um arquivo de suspense ou marcar os dados.)

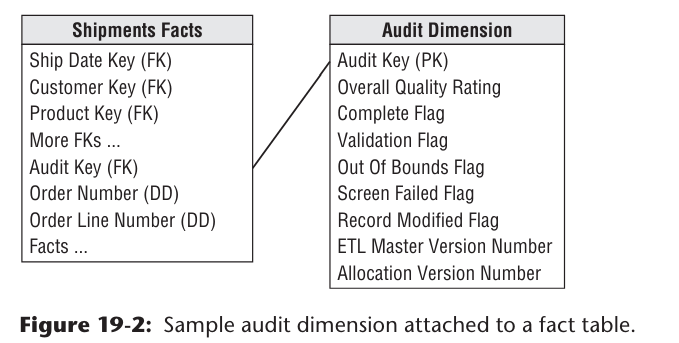
A tabela de fatos de evento de erro também tem uma chave primária de coluna única, mostrada como a chave de evento de erro. Esta chave substituta, como as chaves primárias das tabelas de dimensão, é um simples inteiro atribuído sequencialmente conforme as linhas são adicionadas à tabela de fatos. Esta coluna de chave é necessária nessas situações em que uma explosão enorme de linhas de erro é adicionada à tabela de fatos de evento de erro de uma só vez. Esperançosamente isso não acontecerá com você.

O esquema de evento de erro inclui uma segunda tabela de fatos de detalhes de evento de erro em um grão menor. Cada linha nesta tabela identifica um campo individual em um registro específico que participou de um erro. Assim, um erro de estrutura complexa ou regra de negócios que dispara uma única linha de evento de erro na tabela de fatos de evento de erro de nível superior pode gerar muitas linhas nesta tabela de detalhes de evento de erro de grão inferior. As duas tabelas estão vinculadas pela chave de evento de erro, que é uma chave estrangeira nesta tabela de grão inferior. A tabela de detalhes de evento de erro identifica a tabela, o registro, o campo e a condição de erro precisa. Assim, uma descrição completa de erros complexos de vários campos e vários registros é preservada por essas tabelas.

A tabela de detalhes de evento de erro também pode conter um carimbo de data/hora preciso para fornecer uma descrição completa de eventos de erro de limite agregado onde muitos registros geram uma condição de erro ao longo de um período de tempo. Você deve agora apreciar que cada tela de qualidade tem a responsabilidade de povoar essas tabelas no momento de um erro.

**Subsistema 6: Montador de Dimensão de Auditoria**

A dimensão de auditoria é uma dimensão especial que é montada nos bastidores pelo sistema ETL para cada tabela de fatos, conforme discutimos no Capítulo 6: Gestão de Pedidos. A dimensão de auditoria na Figura 19-2 contém o contexto de metadados no momento em que uma linha específica da tabela de fatos é criada. Você pode dizer que elevamos os metadados a dados reais! Para visualizar como as linhas da dimensão de auditoria são criadas, imagine esta tabela de fatos de remessas sendo atualizada uma vez por dia a partir de um arquivo em lote. Suponha que hoje você tenha uma execução perfeita sem erros sinalizados. Nesse caso, você geraria apenas uma linha de dimensão de auditoria, e ela seria anexada a cada linha de fatos carregada hoje. Todas as categorias, pontuações e números de versão seriam os mesmos.



Agora vamos relaxar a forte suposição de uma execução perfeita. Se você tivesse algumas linhas de fatos cujos dólares de desconto desencadeassem um erro fora dos limites, então mais uma linha de dimensão de auditoria seria necessária para sinalizar essa condição.

**Subsistema 7: Sistema de Desduplicação**

Muitas vezes, as dimensões são derivadas de várias fontes. Esta é uma situação comum para organizações que têm muitos sistemas de origem voltados para o cliente que criam e gerenciam tabelas mestre de clientes separadas. As informações do cliente podem precisar ser mescladas de várias linhas de negócios e fontes externas. Às vezes, os dados podem ser combinados através de valores idênticos em alguma coluna chave. No entanto, mesmo quando ocorre uma correspondência definitiva, outras colunas nos dados podem se contradizer, exigindo uma decisão sobre quais dados devem sobreviver.

Infelizmente, raramente há uma coluna universal que torne a operação de mesclagem fácil. Às vezes, as únicas pistas disponíveis são a semelhança de várias colunas. Os diferentes conjuntos de dados sendo integrados e os dados da tabela de dimensão existente podem precisar ser avaliados em diferentes campos para tentar uma correspondência. Às vezes, uma correspondência pode ser baseada em critérios difusos, como nomes e endereços que podem quase corresponder, exceto por diferenças ortográficas menores.

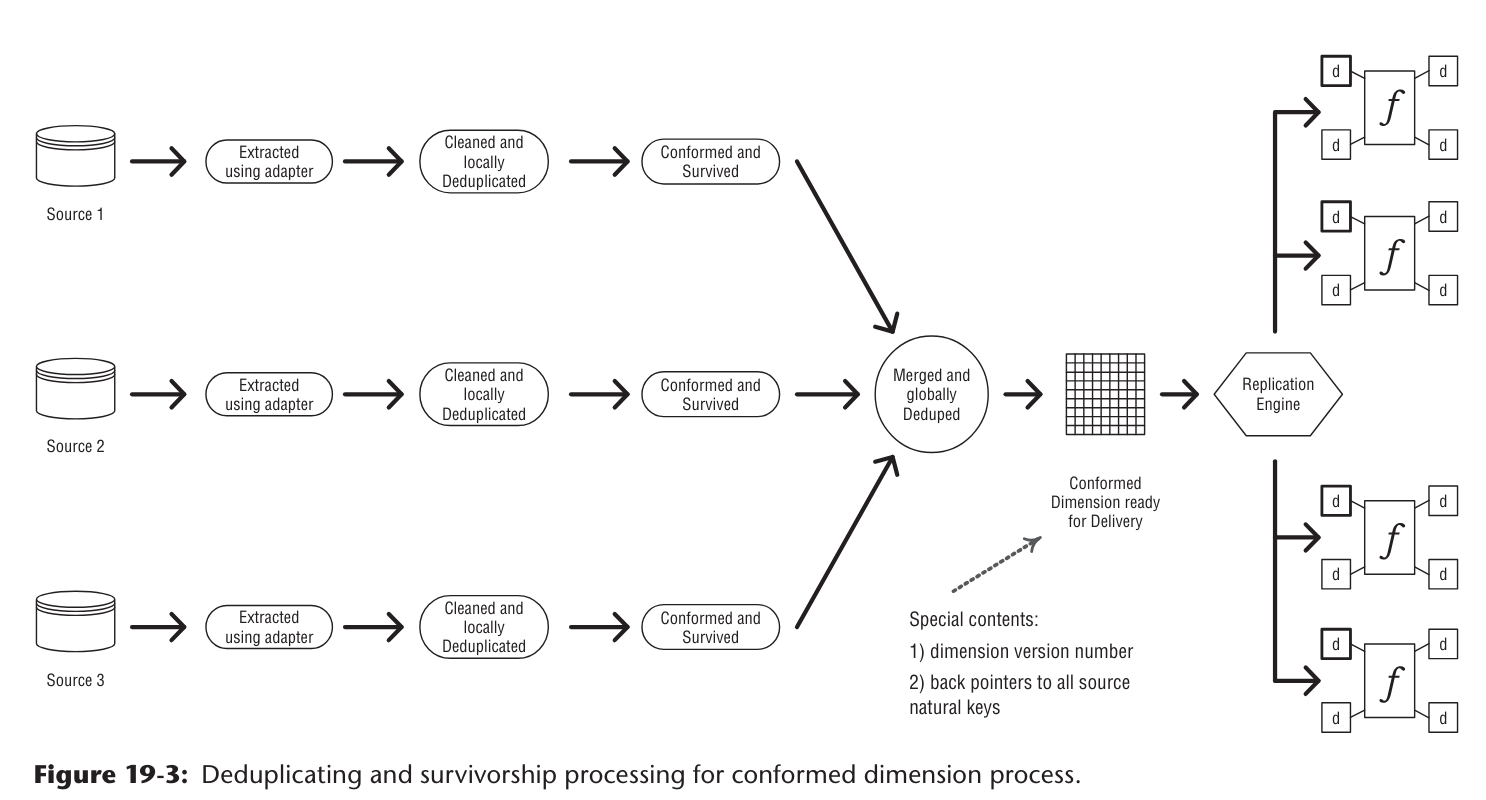
A sobrevivência é o processo de combinar um conjunto de registros correspondidos em uma imagem unificada que combina as colunas de maior qualidade dos registros correspondidos em uma linha conformada. A sobrevivência envolve estabelecer regras de negócios claras que definem a sequência de prioridade para os valores das colunas de todos os sistemas de origem possíveis para permitir a criação de uma única linha com os melhores atributos sobrevividos. Se o design dimensional é alimentado por vários sistemas, você deve manter colunas separadas com referências de volta, como chaves naturais, para todos os sistemas de origem participantes usados para construir a linha.

Há uma variedade de ferramentas de integração de dados e padronização de dados a considerar se você tiver problemas difíceis de desduplicação, correspondência e sobrevivência de dados. Essas ferramentas são bastante maduras e de uso generalizado.

**Subsistema 8: Sistema de Conformidade**

A conformidade consiste em todas as etapas necessárias para alinhar o conteúdo de algumas ou todas as colunas em uma dimensão com colunas em dimensões semelhantes ou idênticas em outras partes do data warehouse. Por exemplo, em uma grande organização, você pode ter tabelas de fatos que capturam faturas e chamadas de serviço ao cliente que utilizam a dimensão do cliente. É altamente provável que os sistemas de origem para faturas e serviço ao cliente tenham bancos de dados de clientes separados. É provável que haja pouca consistência garantida entre as duas fontes de informações do cliente. Os dados dessas duas fontes de clientes precisam ser conformados para fazer com que algumas ou todas as colunas que descrevem o cliente compartilhem os mesmos domínios.

**NOTA** O processo de criação de dimensões conformadas está alinhado com uma abordagem ágil. Para que duas dimensões sejam conformadas, elas devem compartilhar pelo menos um atributo comum com o mesmo nome e mesmo conteúdo. Você pode começar com um único atributo conformado, como Categoria do Cliente, e adicionar sistematicamente essa coluna de maneira não disruptiva às dimensões do cliente em cada um dos processos voltados para o cliente. À medida que você aumenta cada processo voltado para o cliente, você expande a lista de processos que são integrados e podem participar de consultas de drill-across. Você também pode crescer incrementalmente a lista de atributos conformados, como cidade, estado e país. Tudo isso pode ser organizado para se alinhar com uma abordagem de implementação mais ágil.



O subsistema de conformidade é responsável por criar e manter as dimensões conformadas e fatos conformados descritos no Capítulo 4: Inventário. Para realizar isso, os dados de entrada de vários sistemas precisam ser combinados e integrados, para que sejam estruturalmente idênticos, desduplicados, filtrados de dados inválidos e padronizados em termos de linhas de conteúdo em uma imagem conformada. Uma grande parte do processo de conformidade é os processos de desduplicação, correspondência e sobrevivência descritos anteriormente. O fluxo de processo de conformidade que combina os processos de desduplicação e sobrevivência é mostrado na Figura 19-3.

O processo de definição e entrega de dimensões e fatos conformados é descrito posteriormente nos subsistemas 17 (gerente de dimensão) e 18 (provedor de fatos).

**Entregando: Preparar para Apresentação**

A missão primária do sistema ETL é a entrega das tabelas de dimensão e fatos na etapa de entrega. Por essa razão, os subsistemas de entrega são os subsistemas mais cruciais na arquitetura ETL. Embora haja considerável variação nas estruturas de dados de origem e na lógica de limpeza e conformidade, as técnicas de processamento de entrega para preparar as estruturas das tabelas dimensionais são mais definidas e disciplinadas. O uso dessas técnicas é crítico para construir um data warehouse dimensional bem-sucedido que seja confiável, escalável e sustentável.

Muitos desses subsistemas se concentram no processamento de tabelas de dimensão. As tabelas de dimensão são o coração do data warehouse. Elas fornecem o contexto para as tabelas de fatos e, portanto, para todas as medições. Embora as tabelas de dimensão geralmente sejam menores do que as tabelas de fatos, elas são críticas para o sucesso do sistema DW/BI, pois fornecem os pontos de entrada para as tabelas de fatos. O processo de entrega começa com os dados limpos e conformados resultantes dos subsistemas recém descritos. Para muitas dimensões, o plano de carga básico é relativamente simples: você realiza transformações básicas nos dados para construir linhas de dimensão para carregamento na tabela de apresentação de destino. Isso normalmente inclui atribuição de chave substituta, consultas de código para fornecer descrições apropriadas, divisão ou combinação de colunas para apresentar os valores de dados apropriados ou junção de estruturas de tabela em forma normalizada subjacente em dimensões achatadas e desnormalizadas.

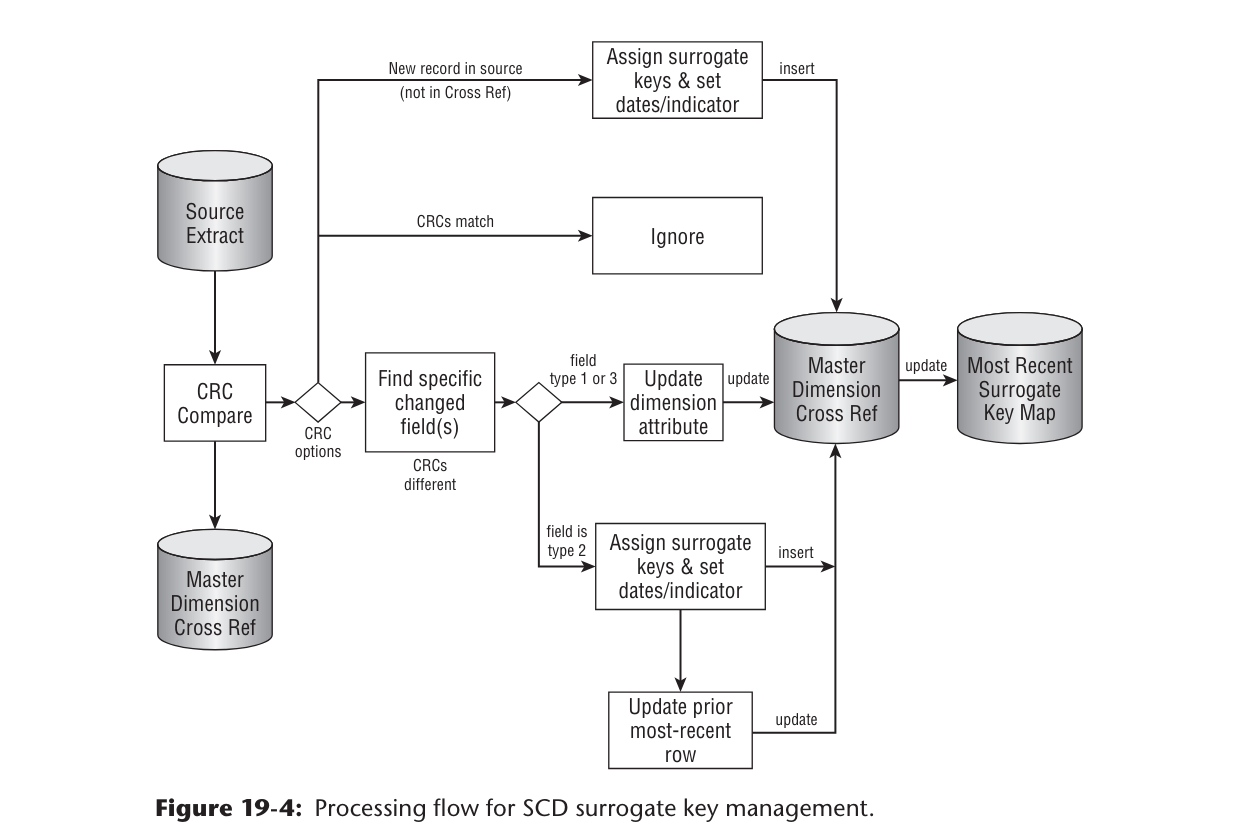
Preparar tabelas de fatos é certamente importante porque as tabelas de fatos contêm as principais medições do negócio que os usuários desejam ver. As tabelas de fatos podem ser grandes e demoradas para carregar. No entanto, preparar tabelas de fatos para apresentação é tipicamente mais direto.

**Subsistema 9: Gerente de Dimensão de Mudança Lenta** (SCD)

Um dos elementos mais importantes da arquitetura ETL é a capacidade de implementar a lógica de dimensão de mudança lenta (SCD). O sistema ETL deve determinar como lidar com um valor de atributo que mudou do valor já armazenado no data warehouse. Se a descrição revisada for determinada como uma atualização legítima e confiável de informações anteriores, a técnica SCD apropriada deve ser aplicada.

Como descrito no Capítulo 5: Aquisição, quando o data warehouse recebe notificação de que uma linha existente em uma dimensão mudou, há três respostas básicas: tipo 1 sobrescrever, tipo 2 adicionar uma nova linha e tipo 3 adicionar uma nova coluna. O gerente de SCD deve lidar sistematicamente com a variação de tempo nas dimensões usando essas três técnicas, bem como outras técnicas de SCD. Além disso, o gerente de SCD deve manter colunas de manutenção apropriadas para mudanças do tipo 2. A Figura 19-4 mostra o fluxo de processamento geral para lidar com o gerenciamento de chave substituta para processar SCDs.

O processo de captura de mudança de dados descrito no subsistema 2 obviamente desempenha um papel importante na apresentação dos dados alterados ao processo de SCD. Assumindo que o processo de captura de mudança de dados tenha efetivamente entregado mudanças apropriadas, o processo de SCD pode tomar as ações apropriadas.



**Tipo 1: Sobrescrever**

A técnica do tipo 1 é uma simples sobrescrita de um ou mais atributos em uma linha de dimensão existente. Você pega os dados revisados do sistema de captura de mudança de dados e sobrescreve o conteúdo da tabela de dimensão. O tipo 1 é apropriado quando corrigir dados ou quando não há necessidade comercial de manter o histórico de valores anteriores. Por exemplo, você pode receber um endereço de cliente corrigido. Neste caso, a sobrescrita é a escolha certa. Observe que se a tabela de dimensão incluir rastreamento de mudança do tipo 2, você deve sobrescrever a coluna afetada em todas as linhas existentes para aquele cliente específico. As atualizações do tipo 1 devem ser propagadas para frente desde as tabelas de staging armazenadas permanentemente mais antigas até todas as tabelas de staging afetadas, então, se alguma delas for usada para recriar as tabelas de carga finais, o efeito da sobrescrita é preservado.

Algumas ferramentas ETL contêm funcionalidade de UPDATE senão INSERT. Esta funcionalidade pode ser conveniente para o desenvolvedor, mas pode ser um assassino de desempenho. Para máximo desempenho, as atualizações de linhas existentes devem ser segregadas das inserções de novas linhas. Se as atualizações do tipo 1 causarem problemas de desempenho, considere desativar o registro do banco de dados ou usar o carregador em massa do DBMS.

As atualizações do tipo 1 invalidam quaisquer agregados construídos sobre a coluna alterada, então o gerente de dimensão (subsistema 17) deve notificar os provedores de fatos afetados (subsistema 18) para descartar e reconstruir os agregados afetados.

**Tipo 2: Adicionar Nova Linha**

O SCD do tipo 2 é a técnica padrão para rastrear com precisão as mudanças nas dimensões e associá-las corretamente com linhas de fatos. Apoiar mudanças do tipo 2 requer um sistema de captura de mudança de dados forte para detectar mudanças assim que ocorrerem. Para atualizações do tipo 2, copie a versão anterior da linha de dimensão e crie uma nova linha de dimensão com uma nova chave substituta. Se não houver uma versão anterior da linha de dimensão, crie uma nova do zero. Em seguida, atualize esta linha com as colunas que mudaram e adicione quaisquer outras colunas que sejam necessárias. Esta é a técnica principal para lidar com mudanças de atributos de dimensão que precisam ser rastreadas ao longo do tempo.

O processo ETL do tipo 2 também deve atualizar a tabela de mapa de chave substituta mais recente, assumindo que a ferramenta ETL não lide automaticamente com isso. Essas pequenas tabelas de duas colunas são de imensa importância ao carregar dados da tabela de fatos. O subsistema 14, o pipeline de chave substituta, apoia esse processo.

Consulte a Figura 19-4 para ver a lógica de consulta e atribuição de chave para lidar com uma linha de dimensão alterada durante o processo de extração. Neste exemplo, o processo de captura de mudança de dados (subsistema 2) usa uma comparação CRC para determinar quais linhas mudaram nos dados de origem desde a última atualização. Se você tiver sorte, você já sabe quais registros de dimensão mudaram e pode omitir essa etapa de comparação CRC. Depois de identificar as linhas que têm mudanças em atributos do tipo 2, você pode gerar uma nova chave substituta a partir da sequência de chaves e atualizar a tabela de mapa de chave substituta.

Quando uma nova linha do tipo 2 é criada, você precisa de pelo menos um par de carimbos de data e hora, bem como um atributo de descrição de mudança opcional. O par de carimbos de data e hora define um intervalo de tempo do início do tempo efetivo até o fim do tempo efetivo quando o conjunto completo de atributos de dimensão é válido. Um tratamento mais sofisticado de uma linha SCD do tipo 2 envolve adicionar cinco colunas de manutenção ETL. Referindo-se à Figura 19-4, isso também requer que o processo ETL do tipo 2 encontre a linha efetiva anterior e faça as atualizações apropriadas nessas colunas de manutenção:

* Data da Mudança (data da mudança como chave estrangeira para a dimensão auxiliar de data)
* Data/Hora Efetiva da Linha (carimbo de data/hora exato da mudança)
* Data/Hora de Fim da Linha (carimbo de data/hora exato da próxima mudança, padrão para 31/12/9999 para a linha de dimensão mais atual)
* Coluna de Motivo da Mudança (atributo opcional)
* Bandeira Atual (atual/expirada)

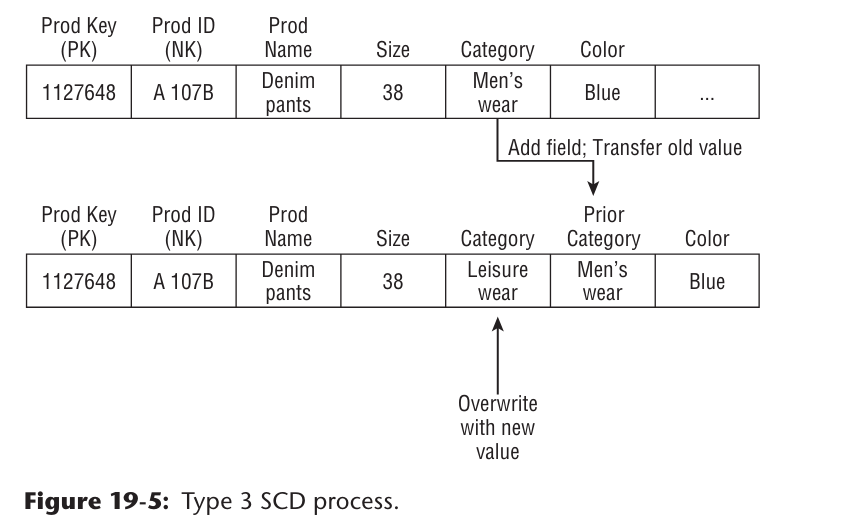
**NOTA** É possível que scripts de back-end sejam executados dentro do banco de dados de transações para modificar dados sem atualizar os respectivos campos de metadados, como o last\_modified\_date. Usar esses campos para os carimbos de data/hora da dimensão pode causar resultados inconsistentes no data warehouse. Sempre use a data do sistema ou a data de corte para derivar os carimbos de data/hora efetivos do tipo 2.

O processo do tipo 2 não altera a história como o processo do tipo 1; portanto, mudanças do tipo 2 não requerem a reconstrução de tabelas agregadas afetadas, desde que a mudança tenha sido feita "hoje" e não retroativamente no tempo.

**NOTA** A Dica de Design Kimball #80 (disponível em [www.kimballgroup.com](http://www.kimballgroup.com/) na aba Ferramentas e Utilitários para este título de livro) fornece orientações detalhadas sobre como adicionar um atributo de código de motivo de mudança de linha às tabelas de dimensão.

**Tipo 3: Adicionar Novo Atributo**

A técnica do tipo 3 é projetada para suportar mudanças "suaves" de atributos que permitem que um usuário se refira tanto ao valor antigo do atributo quanto ao novo. Por exemplo, se uma equipe de vendas for atribuída a uma região de vendas recém-nomeada, pode haver uma necessidade de rastrear a antiga atribuição de região, bem como a nova. A técnica do tipo 3 requer que o sistema ETL altere a tabela de dimensão para adicionar uma nova coluna ao esquema, se essa situação não foi antecipada. Claro, o DBA designado para trabalhar com a equipe ETL provavelmente será responsável por essa mudança. Você então precisa empurrar os valores de coluna existentes para a coluna recém-criada e preencher a coluna original com os novos valores fornecidos ao sistema ETL. A Figura 19-5 mostra como um SCD do tipo 3 é implementado.



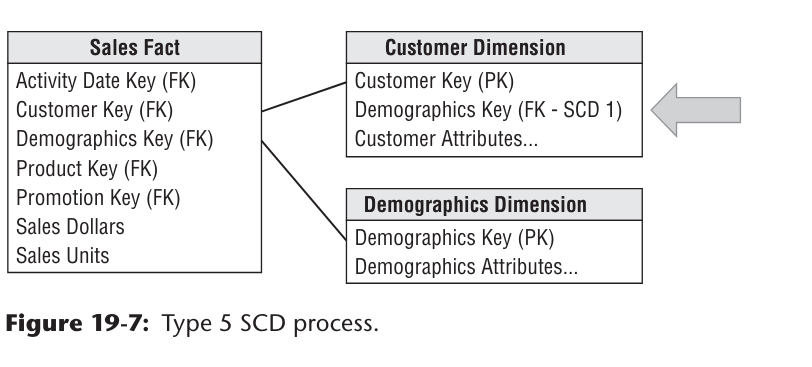
Semelhante ao processo do tipo 1, as atualizações de mudança do tipo 3 invalidam quaisquer agregados construídos sobre a coluna alterada; o gerente de dimensão deve notificar os provedores de fatos afetados, para que eles descartem e reconstruam os agregados afetados.

**Tipo 4: Adicionar Mini-Dimensão**

A técnica do tipo 4 é usada quando um grupo de atributos em uma dimensão muda suficientemente rápido para que sejam divididos em uma mini-dimensão. Esta situação às vezes é chamada de dimensão monstro de mudança rápida. Como o tipo 3, essa situação exige uma mudança de esquema, esperançosamente feita no momento do design. A mini-dimensão requer sua própria chave primária única, e tanto a chave primária da dimensão principal quanto a chave primária da mini-dimensão devem aparecer na tabela de fatos. A Figura 19-6 mostra como um SCD do tipo 4 é implementado.

**Tipo 5: Adicionar Mini-Dimensão e Tipo 1 Outrigger**

A técnica do tipo 5 se baseia na mini-dimensão do tipo 4, também incorporando uma referência do tipo 1 à mini-dimensão na dimensão principal. Isso permite acessar os valores atuais na mini-dimensão diretamente da dimensão base sem vinculação através de uma tabela de fatos. A equipe ETL deve adicionar a referência de chave do tipo 1 na dimensão base e deve sobrescrever essa referência de chave em todas as cópias da dimensão base sempre que o status atual da mini-dimensão mudar ao longo do tempo. A Figura 19-7 mostra como um SCD do tipo 5 é implementado.

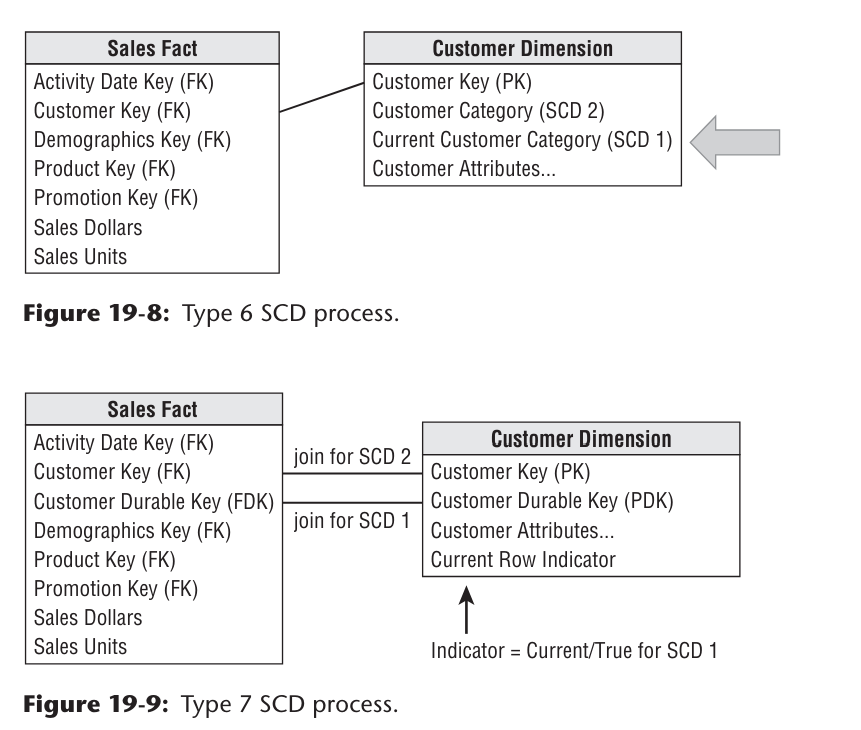


**Tipo 6: Adicionar Atributos do Tipo 1 à Dimensão do Tipo 2**

A técnica do tipo 6 tem um atributo embutido que é um valor alternativo de um atributo normal do tipo 2 na dimensão base. Geralmente, tal atributo é simplesmente uma realidade alternativa do tipo 3, mas neste caso o atributo é sistematicamente sobrescrito sempre que o atributo é atualizado. A Figura 19-8 mostra como um SCD do tipo 6 é implementado.

**Tipo 7: Dimensões Tipo 1 e Tipo 2 Duplas**

A técnica do tipo 7 é uma dimensão tipo 2 normal emparelhada com uma tabela de fatos especialmente construída que tem tanto uma chave estrangeira normal para a dimensão para processamento histórico do tipo 2, quanto também uma chave durável estrangeira (FDK na Figura 19-9) que é usada alternativamente para processamento atual do tipo 1, conectada à chave durável na tabela de dimensão rotulada PDK. A tabela de dimensão também contém um indicador de linha atual que indica se a linha particular é aquela a ser usada para a perspectiva atual do SCD 1. A equipe ETL deve aumentar uma tabela de fatos normalmente construída com essa chave estrangeira durável constante. A Figura 19-9 mostra como um SCD do tipo 7 é implementado.



### Subsistema 10: Gerador de Chaves Surrogadas

Como você deve se lembrar do Capítulo 3: Vendas no Varejo, recomendamos fortemente o uso de chaves surrogadas para todas as tabelas de dimensão. Isso implica que você precisa de um mecanismo robusto para produzir chaves surrogadas no sistema ETL. O gerador de chaves surrogadas deve gerar chaves surrogadas de forma independente para cada dimensão; ele deve ser independente da instância de banco de dados e capaz de servir clientes distribuídos. O objetivo do gerador de chaves surrogadas é gerar uma chave sem significado, tipicamente um inteiro, para servir como chave primária para uma linha de dimensão.

Embora possa ser tentador criar chaves surrogadas por meio de gatilhos de banco de dados, essa técnica pode criar gargalos de desempenho. Se o SGBD for usado para atribuir chaves surrogadas, é preferível que o processo ETL chame diretamente o gerador de sequências do banco de dados. Para melhorar a eficiência, considere fazer com que a ferramenta ETL gere e mantenha as chaves surrogadas. Evite a tentação de concatenar a chave operacional do sistema de origem e um carimbo de data/hora. Embora essa abordagem pareça simples, ela está repleta de problemas e, em última análise, não escalará.

### Subsistema 11: Gerenciador de Hierarquias

É normal que uma dimensão tenha múltiplas estruturas hierárquicas simultâneas e embutidas. Essas múltiplas hierarquias simplesmente coexistem na mesma dimensão como atributos de dimensão. Tudo o que é necessário é que cada atributo seja de valor único na presença da chave primária da dimensão. As hierarquias podem ser fixas ou irregulares. Uma hierarquia de profundidade fixa tem um número consistente de níveis e é simplesmente modelada e populada como atributos de dimensão separados para cada um dos níveis. Hierarquias ligeiramente irregulares, como endereços postais, são geralmente modeladas como uma hierarquia fixa. Hierarquias profundamente irregulares são tipicamente encontradas em estruturas organizacionais que são desbalanceadas e de profundidade indeterminada. O modelo de dados e a solução ETL necessários para suportar essas necessidades requerem o uso de uma tabela de ponte contendo o mapa organizacional.

Estruturas de dados normalizadas ou em floco de neve não são recomendadas para o nível de apresentação. No entanto, o uso de um design normalizado pode ser apropriado na área de preparação ETL para ajudar na manutenção do fluxo de dados ETL para preencher e manter os atributos da hierarquia. O sistema ETL é responsável por impor as regras de negócios para garantir que a hierarquia seja preenchida adequadamente na tabela de dimensão.

### Subsistema 12: Gerenciador de Dimensões Especiais

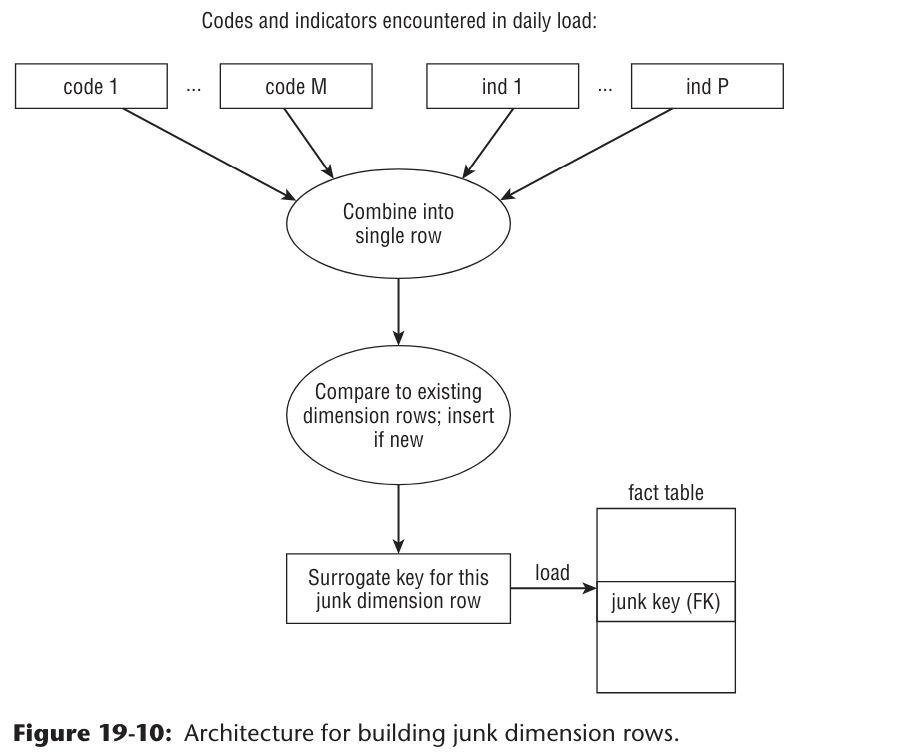
O gerenciador de dimensões especiais é um subsistema abrangente: um espaço reservado na arquitetura ETL para suportar características específicas de design dimensional de uma organização. Alguns sistemas ETL das organizações exigem todas as capacidades discutidas aqui, enquanto outros se preocuparão com poucas dessas técnicas de design:

**Dimensões de Data/Hora**

As dimensões de data e hora são únicas no sentido de que são completamente especificadas no início do projeto de data warehouse, e não têm uma fonte convencional. Isso é aceitável! Tipicamente, essas dimensões são construídas em uma tarde com uma planilha. Mas em um ambiente de empresa global, até mesmo essa dimensão pode ser desafiadora ao considerar múltiplos períodos de relatórios financeiros ou múltiplos calendários culturais.

**Dimensões de Lixo**

Dimensões de lixo são compostas por texto e bandeiras diversas que sobraram na tabela de fatos após a remoção de todos os atributos críticos. Existem duas abordagens para criar dimensões de lixo no sistema ETL. Se o número teórico de linhas na dimensão for fixo e conhecido, a dimensão de lixo pode ser criada antecipadamente. Em outros casos, pode ser necessário criar novas linhas de dimensão de lixo observadas enquanto processa a entrada da linha de fatos. Como ilustrado na Figura 19-10, esse processo requer montar os atributos da dimensão de lixo e compará-los com as linhas existentes da dimensão de lixo para ver se a linha já existe. Caso contrário, uma nova linha de dimensão deve ser montada, uma chave surrogada criada, e a linha carregada na dimensão de lixo durante o processo de carga da tabela de fatos.



**Dimensões Mini**

Como discutimos anteriormente no subsistema 9, dimensões mini são uma técnica usada para rastrear mudanças de atributos dimensionais em uma dimensão grande quando a técnica de tipo 2 é inviável, como uma dimensão de cliente. Do ponto de vista do ETL, a criação da mini-dimensão é semelhante ao processo de dimensão de lixo descrito anteriormente. Novamente, existem duas alternativas: construir todas as combinações válidas antecipadamente ou reconhecer e criar novas combinações conforme necessário. Embora dimensões de lixo sejam geralmente construídas a partir da entrada da tabela de fatos, mini-dimensões são construídas a partir das entradas da tabela de dimensões. O sistema ETL é responsável por manter uma tabela de pesquisa de chaves surrogadas multicolunas para identificar o membro da dimensão base e a linha apropriada da mini-dimensão para suportar o processo do pipeline de chaves surrogadas descrito no Subsistema 14, Pipeline de Chaves Surrogadas. Lembre-se de que dimensões de clientes muito grandes e complexas frequentemente requerem várias mini-dimensões.

**Dimensões Subconjuntos Encolhidos**

Dimensões encolhidas são dimensões conformadas que são um subconjunto de linhas e/ou colunas de uma de suas dimensões base. O fluxo de dados do ETL deve construir dimensões encolhidas conformadas a partir da dimensão base, em vez de de forma independente, para garantir a conformidade. A chave primária da dimensão encolhida, no entanto, deve ser gerada de forma independente; se você tentar usar uma chave de uma linha de dimensão base "exemplar", terá problemas se essa chave for desativada ou substituída.

**Dimensões Estáticas Pequenas**

Poucas dimensões são criadas inteiramente pelo sistema ETL sem uma fonte externa real. Estas são geralmente pequenas dimensões de pesquisa onde um código operacional é traduzido em palavras. Nesses casos, não há realmente processamento ETL. A dimensão de pesquisa é simplesmente criada diretamente pela equipe ETL como uma tabela relacional em sua forma final.

**Dimensões Mantidas pelo Usuário**

Freqüentemente, o data warehouse requer que tabelas de dimensões "mestre" totalmente novas sejam criadas. Essas dimensões não têm um sistema formal de registro; em vez disso, são descrições personalizadas, agrupamentos e hierarquias criadas pelo negócio para fins de relatórios e análise. A equipe ETL muitas vezes acaba com a responsabilidade de gestão dessas dimensões, mas isso geralmente não é bem-sucedido porque a equipe ETL não está ciente das mudanças que ocorrem nesses agrupamentos personalizados, então as dimensões caem em desuso e se tornam ineficazes. O melhor cenário é que o departamento de negócios apropriado concorde em ser o responsável pela manutenção desses atributos. A equipe DW/BI precisa fornecer uma interface de usuário para essa manutenção. Normalmente, isso assume a forma de um aplicativo simples construído usando a ferramenta de programação visual padrão da empresa. O sistema ETL deve adicionar valores de atributos padrão para novas linhas, que o proprietário do usuário precisa atualizar. Se essas linhas forem carregadas no data warehouse antes de serem alteradas, ainda aparecerão em relatórios com qualquer descrição padrão fornecida.

**NOTA:** O processo ETL deve criar uma descrição de atributo de dimensão padrão única que mostre que alguém ainda não fez seu trabalho de gestão de dados. Preferimos um rótulo que concatena a frase Não Atribuído Ainda com o valor da chave surrogada: "Não Atribuído Ainda 157". Dessa forma, múltiplos valores não atribuídos não são agrupados inadvertidamente em relatórios e tabelas agregadas. Isso também ajuda a identificar a linha para correção posterior.

### Subsistema 13: Construtores de Tabelas de Fatos

As tabelas de fatos contêm as medidas de uma organização. Modelos dimensionais são deliberadamente construídos em torno dessas medidas numéricas. O subsistema construtor de tabelas de fatos se concentra nos requisitos arquitetônicos do ETL para construir efetivamente os três principais tipos de tabelas de fatos: transação, snapshot periódico e snapshot acumulativo. Um requisito importante para carregar tabelas de fatos é manter a integridade referencial com as tabelas de dimensão associadas. O pipeline de chaves surrogadas (subsistema 14) é projetado para ajudar a suportar essa necessidade.

**Carregador de Tabelas de Fatos de Transação**

O grão de transação representa um evento de medição definido em um instante particular. Um item de linha em uma fatura é um exemplo de evento de transação. Um evento de scanner em um caixa registradora é outro. Nesses casos, o carimbo de data/hora na tabela de fatos é muito simples. É uma única chave estrangeira de grão diário ou um par consistindo de uma chave estrangeira de grão diário junto com um carimbo de data/hora, dependendo do que o sistema de origem fornece e das análises necessárias. Os fatos nesta tabela de transações devem ser verdadeiros para o grão e devem descrever apenas o que aconteceu naquele instante.

Tabelas de fatos de grão de transação são as maiores e mais detalhadas dos três tipos de tabelas de fatos. O carregador de tabela de fatos de transação recebe dados do sistema de captura de dados alterados e os carrega com as chaves estrangeiras dimensionais adequadas. A adição pura dos registros mais recentes é o caso mais fácil: simplesmente carregar em bloco novas linhas na tabela de fatos. Na maioria dos casos, a tabela de fatos alvo deve ser particionada por tempo para facilitar a administração e acelerar o desempenho da tabela. Uma chave de auditoria, ID sequencial ou coluna de carimbo de data/hora deve ser incluída para permitir o backup ou reinício da tarefa de carga.

A adição de dados atrasados é mais difícil, exigindo capacidades de processamento adicionais descritas no subsistema 16. No caso de ser necessário atualizar linhas existentes, este processo deve ser realizado em duas fases. A primeira etapa é inserir as linhas corrigidas sem sobrescrever ou excluir as linhas originais, e então excluir as linhas antigas em uma segunda etapa. Usando uma chave surrogada única atribuída sequencialmente para a tabela de fatos, é possível realizar as duas etapas de inserção seguidas de exclusão.

**Carregador de Tabelas de Fatos de Snapshot Periódico**

O grão de snapshot periódico representa uma medição regular e repetitiva ou um conjunto de medições, como um extrato mensal de conta bancária. Esta tabela de fatos também tem uma única coluna de data, representando o período geral. Os fatos nesta tabela de snapshot periódico devem ser verdadeiros para o grão e devem descrever apenas medidas apropriadas ao período definido. Snapshots periódicos são um tipo comum de tabela de fatos e são frequentemente usados para saldos de contas, relatórios financeiros mensais e saldos de inventário. A periodicidade de um snapshot periódico é tipicamente diária, semanal ou mensal.

Snapshots periódicos têm características de carregamento semelhantes às tabelas de fatos de grão de transação. O mesmo processamento se aplica para inserções e atualizações. Supondo que os dados sejam prontamente entregues ao sistema ETL, todos os registros para cada carga periódica podem se agrupar na partição de tempo mais recente. Tradicionalmente, snapshots periódicos têm sido carregados em massa no final do período apropriado.

Por exemplo, uma empresa de cartão de crédito pode carregar uma tabela de snapshot de conta mensal com os saldos em vigor no final do mês. Mais frequentemente, as organizações populam um snapshot periódico quente e rolante. Além das linhas carregadas no final de cada mês, há linhas especiais carregadas com os saldos mais recentes em vigor a partir do dia anterior. À medida que o mês progride, as linhas do mês atual são continuamente atualizadas com as informações mais recentes e continuam dessa maneira rolando ao longo do mês. Observe que o snapshot quente e rolante pode às vezes ser difícil de implementar se as regras de negócios para calcular os saldos no final do período forem complexas. Frequentemente, esses cálculos complexos dependem de outros processamentos periódicos fora do data warehouse, e não há informações suficientes disponíveis para o sistema ETL realizar esses cálculos complexos com mais frequência.

**Carregador de Tabelas de Fatos de Snapshot Acumulativo**

O grão de snapshot acumulativo representa o status evolutivo atual de um processo que tem um início e um fim definidos. Normalmente, esses processos são de curta duração e, portanto, não se prestam ao snapshot periódico. O processamento de pedidos é o exemplo clássico de um snapshot acumulativo. O pedido é feito, enviado e pago dentro de um período de relatório. O grão de transação fornece detalhes demais separados em linhas individuais da tabela de fatos, e o snapshot periódico é simplesmente a maneira errada de relatar esses dados.

O design e a administração do snapshot acumulativo são bastante diferentes dos dois primeiros tipos de tabelas de fatos. Todas as tabelas de fatos de snapshot acumulativo têm um conjunto de datas que descrevem o fluxo de trabalho típico do processo. Por exemplo, um pedido pode ter uma data de pedido, data de envio real, data de entrega, data de pagamento final e data de devolução. Neste exemplo, essas cinco datas aparecem como cinco chaves estrangeiras de valor de data separadas. Quando a linha do pedido é criada pela primeira vez, a primeira dessas datas está bem definida, mas talvez nenhuma das outras tenha ainda acontecido. Esta mesma linha de fatos é subsequentemente revisitada à medida que o pedido avança pelo pipeline de pedidos. Cada vez que algo acontece, a linha de fatos de snapshot acumulativo é modificada de forma destrutiva. As chaves estrangeiras de data são sobrescritas, e vários fatos são atualizados. Frequentemente, a primeira data permanece inviolável porque descreve quando a linha foi criada, mas todas as outras datas podem muito bem ser sobrescritas, às vezes mais de uma vez.

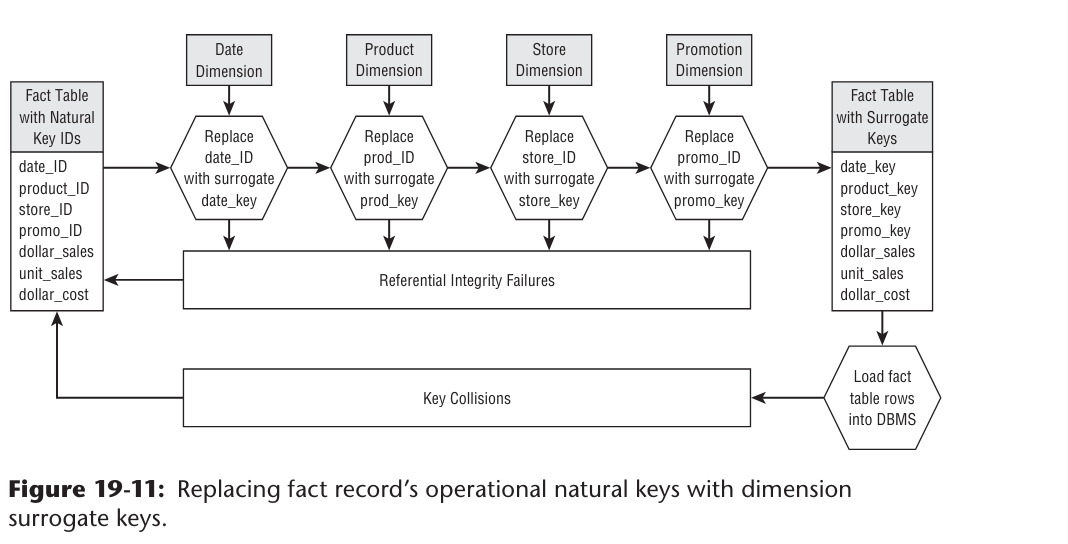
Muitos SGBDs utilizam comprimentos variáveis de linhas. Atualizações repetidas às linhas de fatos de snapshot acumulativo podem fazer com que as linhas cresçam devido a esses comprimentos variáveis de linhas, afetando a residência dos blocos de disco. Pode valer a pena ocasionalmente soltar e recarregar linhas após a atividade de atualização para melhorar o desempenho.

Uma tabela de fatos de snapshot acumulativo é uma maneira eficaz de representar processos finitos com inícios e términos bem definidos. No entanto, o snapshot acumulativo, por definição, é a visão mais recente. Frequentemente, faz sentido utilizar todos os três tipos de tabelas de fatos para atender a várias necessidades. A história periódica pode ser capturada com extratos periódicos, e todos os detalhes infinitos envolvidos no processo podem ser capturados em uma tabela de fatos de grão de transação associada. A presença de muitas situações que violam cenários padrão ou envolvem repetidas passagens pelo processo proibiria o uso de um snapshot acumulativo.

### Subsistema 14: Pipeline de Chaves Surrogadas

Todo sistema ETL deve incluir uma etapa para substituir as chaves naturais operacionais na linha de tabela de fatos de entrada pelas chaves surrogadas de dimensão apropriadas. A integridade referencial (RI) significa que para cada chave estrangeira na tabela de fatos, deve existir uma entrada na tabela de dimensão correspondente. Se houver uma linha em uma tabela de fatos de vendas para a chave surrogada do produto 323442, você precisa ter uma linha na tabela de dimensão do produto com a mesma chave, ou você não saberá o que vendeu. Você tem uma venda para o que parece ser um produto inexistente. Pior ainda, sem a chave do produto na dimensão, um usuário de negócios pode facilmente construir uma consulta que omitirá essa venda sem perceber.

O processo de pesquisa de chaves deve resultar em uma correspondência para cada chave natural de entrada ou um valor padrão. No caso de haver uma falha de integridade referencial não resolvida durante o processo de pesquisa, você precisa alimentar essas falhas de volta para o processo ETL responsável para resolução, conforme mostrado na Figura 19-11. Da mesma forma, o processo ETL precisa resolver quaisquer colisões de chave que possam ser encontradas durante o processo de pesquisa de chaves.



Após os dados da tabela de fatos terem sido processados e antes de carregar na camada de apresentação, uma pesquisa de chave surrogada precisa ocorrer para substituir as chaves naturais operacionais no registro da tabela de fatos pela chave surrogada atual adequada. Para preservar a integridade referencial, sempre complete a atualização das tabelas de dimensão primeiro. Dessa forma, as tabelas de dimensão são sempre a fonte legítima de chaves primárias que você deve substituir na tabela de fatos (consulte a Figura 19-11).

A abordagem mais direta é usar a própria tabela de dimensão como a fonte para o valor mais atual da chave surrogada correspondente a cada chave natural. Cada vez que você precisar da chave surrogada atual, consulte todas as linhas na dimensão com a chave natural igual ao valor desejado e, em seguida, selecione a chave surrogada que se alinha com o contexto histórico da linha de fatos usando o indicador de linha atual ou as datas de início e término de efeito. Ambientes de hardware atuais oferecem memória endereçável quase ilimitada, tornando essa abordagem prática.

Durante o processamento, cada chave natural no registro de fato de entrada é substituída pela chave surrogada atual correta. Não mantenha a chave natural na linha de fatos—a tabela de fatos deve conter apenas a chave surrogada. Não escreva os dados de entrada no disco até que todas as linhas de fatos tenham passado por todas as etapas de processamento. Se possível, todas as tabelas de dimensão necessárias devem ser fixadas na memória, para que possam ser acessadas aleatoriamente à medida que cada registro de entrada apresenta suas chaves naturais.

Como ilustrado na parte inferior da Figura 19-11, o pipeline de chaves surrogadas precisa lidar com colisões de chave no caso de você tentar carregar uma linha duplicada. Este é um exemplo de um problema de qualidade de dados apropriado para uma triagem de qualidade de dados de estrutura tradicional, conforme discutido no subsistema 4. No caso de uma colisão de chave ser reconhecida, o processo do pipeline de chaves surrogadas precisa escolher parar o processo, enviar os dados ofensivos para suspensão ou aplicar as regras de negócios apropriadas para determinar se é possível corrigir o problema, carregar a linha e escrever uma linha explicativa no esquema de eventos de erro.

Observe que um processo ligeiramente diferente é necessário para realizar pesquisas de chaves surrogadas se você precisar recarregar a história ou se tiver muitas linhas de fatos que chegam atrasadas, porque você não quer mapear o valor mais atual para um evento histórico. Nesse caso, você precisa criar lógica para encontrar a chave surrogada que se aplicava no momento em que o registro de fatos foi gerado. Isso significa encontrar a chave surrogada onde a data da transação de fatos está entre a data de início de efeito da chave e a data de término.

Quando as chaves naturais da tabela de fatos foram substituídas por chaves surrogadas, a linha de fatos está pronta para ser carregada. As chaves na linha da tabela de fatos foram escolhidas para serem chaves estrangeiras adequadas, e a tabela de fatos está garantida para ter integridade referencial em relação às tabelas de dimensão.

### Subsistema 15: Construtor de Tabelas de Ponte de Dimensão Multivalorada

Às vezes, uma tabela de fatos deve suportar uma dimensão que assume múltiplos valores na granularidade mais baixa da tabela de fatos, como descrito no Capítulo 8: Gestão de Relacionamento com o Cliente. Se o grão da tabela de fatos não puder ser alterado para suportar diretamente essa dimensão, então a dimensão multivalorada deve ser vinculada à tabela de fatos por meio de uma tabela de ponte. Tabelas de ponte são comuns na indústria de saúde, em ambientes de comissões de vendas e para suportar hierarquias de profundidade variável, conforme discutido no subsistema 11.

O desafio para a equipe de ETL é construir e manter a tabela de ponte. À medida que relações multivaloradas com a linha de fatos são encontradas, o sistema ETL tem a escolha de tornar cada conjunto de observações um grupo único ou reutilizar grupos quando um conjunto idêntico de observações ocorrer. Infelizmente, não há uma resposta simples para a escolha certa. No caso da dimensão multivalorada ter atributos de tipo 2, a tabela de ponte também deve ser variante no tempo, como o conjunto de diagnósticos variantes no tempo de um paciente.

Uma das construções de tabela de ponte apresentadas no Capítulo 10: Serviços Financeiros foi a inclusão de um fator de ponderação para suportar relatórios devidamente ponderados a partir da tabela de ponte. Em muitos casos, o fator de ponderação é um fator de alocação familiar, mas em outros casos, a identificação do fator de ponderação apropriado pode ser problemática porque pode não haver uma base racional para atribuir o fator de ponderação.

### Subsistema 16: Manipulador de Dados Atrasados

Data warehouses geralmente são construídos em torno da suposição ideal de que a atividade medida (registros de fatos) chega ao data warehouse ao mesmo tempo que o contexto da atividade (registros de dimensão). Quando você tem tanto os registros de fatos quanto as linhas de dimensão contemporâneas corretas, você tem o luxo de primeiro manter as chaves da dimensão e depois usar essas chaves atualizadas nas linhas de fatos que os acompanham. No entanto, por uma variedade de razões, o sistema ETL pode precisar processar dados de fatos ou dimensões atrasados.

Em alguns ambientes, pode ser necessário fazer modificações especiais nos procedimentos de processamento padrão para lidar com fatos atrasados, ou seja, registros de fatos que chegam ao data warehouse muito atrasados. Esta é uma situação complicada porque você precisa buscar na história para decidir quais chaves de dimensão estavam em vigor quando a atividade ocorreu. Além disso, pode ser necessário ajustar quaisquer saldos semi-aditivos em linhas de fatos subsequentes. Em um ambiente fortemente regulamentado, também é necessário interagir com o subsistema de conformidade porque você está prestes a alterar a história.

Dimensões atrasadas ocorrem quando a medição da atividade (registro de fatos) chega ao data warehouse sem seu contexto completo. Em outras palavras, os status das dimensões anexadas à medição da atividade são ambíguos ou desconhecidos por algum período de tempo. Se você estiver vivendo no ciclo de atualização em lote convencional de um ou mais dias de latência, você pode geralmente esperar que as dimensões sejam relatadas. Por exemplo, a identificação do novo cliente pode vir em um feed separado atrasado por várias horas; você pode simplesmente esperar até que a dependência seja resolvida.

Mas em muitas situações, especialmente em ambientes em tempo real, esse atraso não é aceitável. Você não pode suspender as linhas e esperar que as atualizações da dimensão ocorram; os requisitos de negócios exigem que você torne a linha de fatos visível antes de conhecer o contexto dimensional. O sistema ETL precisa de capacidades adicionais para suportar esse requisito. Usando cliente como a dimensão problemática, o sistema ETL precisa suportar duas situações. A primeira é suportar atualizações de dimensão de tipo 2 atrasadas. Nesta situação, você precisa adicionar a linha de cliente revisada à dimensão com uma nova chave surrogada e, em seguida, modificar destrutivamente qualquer chave estrangeira de linha de fatos subsequente para a tabela de clientes. As datas de vigência para as linhas de dimensão afetadas também precisam ser redefinidas. Além disso, você precisa fazer uma varredura na dimensão para ver se houve linhas de tipo 2 subsequentes para este cliente e alterar esta coluna em qualquer linha afetada.

A segunda situação ocorre quando você recebe uma linha de fatos com o que parece ser uma chave natural de cliente válida, mas você ainda não carregou esse cliente na dimensão de clientes. Seria possível carregar essa linha apontando para uma linha padrão na tabela de dimensão. Esta abordagem tem o mesmo efeito colateral desagradável discutido anteriormente de exigir atualizações destrutivas das chaves estrangeiras das linhas de fatos quando as atualizações da dimensão são finalmente processadas. Alternativamente, se você acreditar que o cliente é válido, mas ainda não processado, você deve atribuir uma nova chave surrogada de cliente com um conjunto de valores de atributos fictícios em uma nova linha de dimensão de cliente. Você então retorna a esta linha de dimensão fictícia em um momento posterior e faz alterações de sobrescrita de tipo 1 em seus atributos quando obtiver informações completas sobre o novo cliente. Pelo menos este passo evita a alteração destrutiva de quaisquer chaves da tabela de fatos.

Não há como evitar um breve período provisório em que as dimensões estão "não exatamente corretas". Mas essas etapas de manutenção podem minimizar o impacto das atualizações inevitáveis das chaves e de outras colunas.

### Subsistema 17: Sistema de Gerenciamento de Dimensões

O gerente de dimensões é uma autoridade centralizada que prepara e publica dimensões conformadas para a comunidade de data warehouse. Uma dimensão conformada é, por necessidade, um recurso gerenciado centralmente: Cada dimensão conformada deve ter uma única fonte consistente. É responsabilidade do gerente de dimensões administrar e publicar a(s) dimensão(ões) conformada(s) pelas quais ele é responsável. Pode haver múltiplos gerentes de dimensão em uma organização, cada um responsável por uma dimensão. As responsabilidades do gerente de dimensões incluem o seguinte processamento ETL:

* Implementar os rótulos descritivos comuns acordados pelos administradores de dados e partes interessadas durante o design da dimensão.
* Adicionar novas linhas à dimensão conformada para novos dados de origem, gerando novas chaves surrogadas.
* Adicionar novas linhas para mudanças de tipo 2 em entradas de dimensão existentes, gerando novas chaves surrogadas.
* Modificar linhas no local para mudanças de tipo 1 e mudanças de tipo 3, sem alterar as chaves surrogadas.
* Atualizar o número da versão da dimensão se quaisquer mudanças de tipo 1 ou tipo 3 forem feitas.
* Replicar a dimensão revisada simultaneamente para todos os provedores de tabelas de fatos.

É mais fácil gerenciar dimensões conformadas em um único espaço de tabelas SGBD em uma única máquina, porque há apenas uma cópia da tabela de dimensão. No entanto, gerenciar dimensões conformadas torna-se mais difícil em múltiplos espaços de tabelas, múltiplos SGBDs ou ambientes distribuídos em múltiplas máquinas. Nesses casos, o gerente de dimensões deve gerenciar cuidadosamente a liberação simultânea de novas versões da dimensão para cada provedor de fatos. Cada dimensão conformada deve ter uma coluna de número de versão em cada linha que é sobrescrita em cada linha sempre que o gerente de dimensões libera a dimensão. Este número de versão deve ser utilizado para suportar quaisquer consultas de drill-across para garantir que a mesma versão da dimensão está sendo utilizada.

### Subsistema 18: Sistema Provedor de Fatos

O provedor de fatos é responsável por receber dimensões conformadas dos gerentes de dimensões. O provedor de fatos possui a administração de uma ou mais tabelas de fatos e é responsável por sua criação, manutenção e uso. Se tabelas de fatos forem usadas em quaisquer aplicações de drill-across, então, por definição, o provedor de fatos deve estar usando dimensões conformadas fornecidas pelo gerente de dimensões. As responsabilidades do provedor de fatos são mais complexas e incluem:

* Receber ou baixar a dimensão replicada do gerente de dimensões.
* Em um ambiente em que a dimensão não pode ser simplesmente replicada, mas deve ser atualizada localmente, o provedor de fatos deve processar registros de dimensão marcados como novos e atuais para atualizar mapas de chaves atuais no pipeline de chaves surrogadas e também processar quaisquer registros de dimensão marcados como novos, mas pós-datados.
* Adicionar todas as novas linhas às tabelas de fatos após substituir suas chaves naturais pelas chaves surrogadas corretas.
* Modificar linhas em todas as tabelas de fatos para correção de erros, snapshots acumulativos e mudanças de dimensões atrasadas.
* Remover agregados que se tornaram inválidos.
* Recalcular agregados afetados. Se a nova versão de uma dimensão não alterar o número da versão, os agregados devem ser estendidos para lidar apenas com dados de fatos recém-carregados. Se o número da versão da dimensão tiver mudado, o agregado histórico completo pode precisar ser recalculado.
* Garantir a qualidade de todas as tabelas de fatos base e agregadas. Estar satisfeito de que as tabelas de agregados são calculadas corretamente.
* Colocar as tabelas de fatos e dimensões atualizadas online.
* Informar os usuários que o banco de dados foi atualizado. Dizer-lhes se grandes mudanças foram feitas, incluindo mudanças de versão da dimensão, registros pós-datados sendo adicionados e mudanças em agregados históricos.

### Subsistema 19: Construtor de Agregados

Os agregados são a maneira mais dramática de afetar o desempenho em um ambiente de data warehouse grande. As agregações são como índices; são estruturas de dados específicas criadas para melhorar o desempenho. Os agregados podem ter um impacto significativo no desempenho. O sistema ETL precisa construir e usar agregados efetivamente sem causar distrações significativas ou consumir recursos extraordinários e ciclos de processamento.

Você deve evitar arquiteturas em que a navegação de agregados está embutida na ferramenta de consulta proprietária. Do ponto de vista do ETL, o construtor de agregados precisa popular e manter linhas de tabelas de fatos agregadas e tabelas de dimensões encolhidas onde necessário para tabelas de fatos agregadas. A estratégia de atualização mais rápida é incremental, mas uma grande mudança em um atributo de dimensão pode exigir a exclusão e reconstrução do agregado. Em alguns ambientes, pode ser mais rápido despejar dados fora do SGBD e construir agregados com uma utilidade de ordenação em vez de construir os agregados dentro do SGBD. Fatos numéricos aditivos podem ser agregados facilmente no momento da extração calculando linhas de quebra em um dos pacotes de ordenação. Os agregados devem sempre ser consistentes com os dados base atômicos. O provedor de fatos (subsistema 18) é responsável por retirar os agregados de linha quando eles não são consistentes com os dados base.

O feedback dos usuários sobre as consultas que executam lentamente é uma entrada crítica para projetar agregações. Embora você possa depender de feedback informal até certo ponto, um registro de consultas frequentemente tentadas e de execução lenta deve ser capturado. Você também deve tentar identificar as consultas de execução lenta inexistentes que nunca chegaram ao registro porque nunca são concluídas ou nem são tentadas devido a desafios conhecidos de desempenho.

### Subsistema 20: Construtor de Cubos OLAP

Servidores OLAP apresentam dados dimensionais de maneira intuitiva, permitindo que uma gama de usuários analíticos fatie e corte dados. OLAP é um irmão dos esquemas de estrela dimensionais no banco de dados relacional, com inteligência sobre relacionamentos e cálculos definidos no servidor que permitem um desempenho de consulta mais rápido e análises mais interessantes a partir de uma ampla gama de ferramentas de consulta. Não pense em um servidor OLAP como um concorrente de um data warehouse relacional, mas sim como uma extensão. Deixe o banco de dados relacional fazer o que ele faz melhor: fornecer armazenamento e gerenciamento.

O esquema dimensional relacional deve ser visto como a base para cubos OLAP se você optar por incluí-los em sua arquitetura. O processo de alimentar dados do esquema dimensional é uma parte integral do sistema ETL; os esquemas relacionais são a melhor e preferida fonte para cubos OLAP. Como muitos sistemas OLAP não abordam diretamente a integridade referencial ou a limpeza de dados, a arquitetura preferida é carregar cubos OLAP após a conclusão dos processos ETL convencionais. Note que algumas ferramentas OLAP são mais sensíveis a hierarquias do que esquemas relacionais. É importante impor fortemente a integridade das hierarquias dentro das dimensões antes de carregar um cubo OLAP. SCDs de tipo 2 se encaixam bem em um sistema OLAP porque uma nova chave surrogada é tratada como um novo membro. SCDs de tipo 1 que reescrevem a história não se encaixam bem em OLAP. Sobrescritas de um valor de atributo podem fazer com que todos os cubos usando essa dimensão sejam reprocessados em segundo plano, fiquem corrompidos ou sejam descartados. Leia esta última frase novamente.

### Subsistema 21: Gerenciador de Propagação de Dados

O gerenciador de propagação de dados é responsável pelos processos ETL necessários para apresentar dados corporativos conformados e integrados do servidor de apresentação do data warehouse para outros ambientes para fins especiais. Muitas organizações precisam extrair dados da camada de apresentação para compartilhar com parceiros de negócios, clientes e/ou fornecedores para fins estratégicos. Da mesma forma, algumas organizações são obrigadas a enviar dados para várias organizações governamentais para fins de reembolso, como organizações de saúde que participam do programa Medicare. Muitas organizações adquiriram aplicativos analíticos de pacote. Normalmente, esses aplicativos não podem ser apontados diretamente para as tabelas de data warehouse existentes, então os dados precisam ser extraídos da camada de apresentação e carregados nas estruturas de dados proprietárias exigidas pelos aplicativos analíticos. Finalmente, a maioria das ferramentas de mineração de dados não funciona diretamente contra o servidor de apresentação. Elas precisam de dados extraídos do data warehouse e alimentados na ferramenta de mineração de dados em um formato específico.

Todas as situações descritas anteriormente requerem extração do servidor de apresentação DW/BI, possivelmente alguma transformação leve e carregamento em um formato de destino — em outras palavras, ETL. A propagação de dados deve ser considerada uma parte do sistema ETL; as ferramentas ETL devem ser aproveitadas para fornecer essa capacidade. O que é diferente nesta situação é que os requisitos do destino não são negociáveis; você deve fornecer os dados conforme especificado pelo destino.

### Gerenciamento do Ambiente ETL

Um ambiente DW/BI pode ter um ótimo modelo dimensional, aplicativos BI bem implantados e forte patrocínio da gestão. Mas não pode ser um sucesso até que possa ser confiável como uma fonte dependente para a tomada de decisões de negócios. Um dos objetivos do sistema DW/BI é construir uma reputação de fornecer dados oportunos, consistentes e confiáveis para capacitar os negócios. Para alcançar este objetivo, o sistema ETL deve trabalhar constantemente para cumprir três critérios:

* **Confiabilidade:** Os processos ETL devem ser executados consistentemente. Eles devem ser concluídos para fornecer dados em uma base oportuna que seja confiável em qualquer nível de detalhe.
* **Disponibilidade:** O data warehouse deve cumprir seus acordos de nível de serviço (SLAs). O warehouse deve estar disponível conforme prometido.
* **Gerenciabilidade:** Um data warehouse bem-sucedido nunca está concluído. Ele cresce e muda constantemente junto com o negócio. Os processos ETL precisam evoluir graciosamente também.

Os subsistemas de gerenciamento ETL são os principais componentes da arquitetura para ajudar a alcançar os objetivos de confiabilidade, disponibilidade e gerenciabilidade. Operar e manter um data warehouse de maneira profissional não é muito diferente de qualquer outra operação de sistemas: Siga as melhores práticas padrão, planeje para desastres e pratique. A maioria dos subsistemas de gerenciamento necessários que se seguem pode ser familiar para você.

### Subsistema 22: Agendador de Tarefas

Todo data warehouse empresarial deve ter um agendador de ETL robusto. Todo o processo ETL deve ser gerenciado, na medida do possível, por meio de um único ambiente de controle de tarefas baseado em metadados. Os principais fornecedores de ferramentas ETL incluem capacidades de agendamento em seus ambientes. Se você optar por não usar o agendador incluído na ferramenta ETL ou não usar uma ferramenta ETL, você precisa utilizar o agendamento de produção existente ou talvez codificar manualmente as tarefas ETL para execução.

O agendamento é muito mais do que apenas lançar tarefas em um cronograma. O agendador precisa estar ciente e controlar as relações e dependências entre as tarefas ETL. Ele precisa reconhecer quando um arquivo ou tabela está pronto para ser processado. Se a organização estiver processando em tempo real, você precisa de um agendador que suporte a arquitetura de tempo real selecionada. O processo de controle de tarefas também deve capturar metadados sobre o progresso e estatísticas do processo ETL durante sua execução. Finalmente, o agendador deve suportar um processo totalmente automatizado, incluindo notificar o sistema de escalonamento de problemas no caso de qualquer situação que requeira resolução.

A infraestrutura para gerenciar isso pode ser tão básica (e intensiva em mão de obra) quanto um conjunto de procedimentos armazenados SQL, ou tão sofisticada quanto uma ferramenta integrada projetada para gerenciar e orquestrar processos de extração e carga de dados multiplataforma. Se você usar uma ferramenta ETL, ela deve fornecer essa capacidade. Em qualquer caso, você precisa configurar um ambiente para criar, gerenciar e monitorar o fluxo de trabalho ETL.

Os serviços de controle de tarefas necessários incluem:

* **Definição de Tarefas:** O primeiro passo na criação de um processo de operações é ter alguma maneira de definir uma série de etapas como uma tarefa e especificar alguma relação entre tarefas. É aqui que o fluxo de execução do processo ETL é escrito. Em muitos casos, se a carga de uma determinada tabela falhar, isso pode impactar sua capacidade de carregar tabelas que dependem dela. Por exemplo, se a tabela de clientes não for atualizada corretamente, carregar fatos de vendas para novos clientes que não chegaram à tabela de clientes é arriscado. Em alguns bancos de dados, isso é impossível.
* **Agendamento de Tarefas:** No mínimo, o ambiente precisa fornecer capacidades padrão, como agendamento baseado em tempo e evento. Processos ETL são frequentemente baseados em algum evento de sistema upstream, como a conclusão bem-sucedida do fechamento do livro geral ou a aplicação bem-sucedida de ajustes de vendas às figuras de vendas de ontem. Isso inclui a capacidade de monitorar flags de banco de dados, verificar a existência de arquivos e comparar datas de criação.
* **Captura de Metadados:** Nenhuma pessoa responsável por sistemas toleraria um sistema de agendamento como uma caixa preta. As pessoas responsáveis por executar as cargas demandarão um sistema de monitoramento de fluxo de trabalho (subsistema 27) para entender o que está acontecendo. O agendador de tarefas precisa capturar informações sobre em que etapa a carga está, quando começou e quanto tempo levou. Em um sistema ETL feito à mão, isso pode ser realizado fazendo com que cada etapa escreva em um arquivo de log. A ferramenta ETL deve capturar esses dados sempre que um processo ETL for executado.
* **Registro:** Isso significa coletar informações sobre todo o processo ETL, não apenas o que está acontecendo no momento. Informações de registro suportam a recuperação e reinício de um processo em caso de erros durante a execução da tarefa. Registrar em arquivos de texto é o nível mínimo aceitável. Preferimos um sistema que registre em um banco de dados porque a estrutura facilita a criação de gráficos e relatórios. Também torna possível criar estudos de séries temporais para ajudar a analisar e otimizar o processo de carga.
* **Notificação:** Após o processo ETL ter sido desenvolvido e implantado, ele deve ser executado de forma autônoma. Ele deve funcionar sem intervenção humana, sem falhas. Se um problema ocorrer, o sistema de controle precisa se integrar ao sistema de escalonamento de problemas (subsistema 30).

**NOTA:** Alguém precisa saber se algo imprevisto aconteceu durante a carga, especialmente se uma resposta for crítica para continuar o processo.

### Subsistema 23: Sistema de Backup

O data warehouse está sujeito aos mesmos riscos que qualquer outro sistema de computador. Discos rígidos falharão, fontes de energia se desligarão e sistemas de sprinklers se ativarão acidentalmente. Além desses riscos, o warehouse também precisa manter mais dados por períodos mais longos do que os sistemas operacionais. Embora tipicamente não seja gerenciado pela equipe ETL, o processo de backup e recuperação é frequentemente projetado como parte do sistema ETL. Seu objetivo é permitir que o data warehouse volte ao trabalho após uma falha. Isso inclui fazer backup dos dados de preparação intermediária necessários para reiniciar tarefas ETL falhadas. O processo de arquivamento e recuperação é projetado para permitir o acesso do usuário a dados mais antigos que foram movidos do warehouse principal para uma mídia de menor custo, geralmente de desempenho inferior.

**Backup**

Mesmo que você tenha um sistema totalmente redundante com uma fonte de energia universal, discos totalmente RAIDados e processadores paralelos com failover, alguma crise de sistema eventualmente ocorrerá. Mesmo com hardware perfeito, alguém sempre pode excluir a tabela (ou banco de dados) errada. Correndo o risco de afirmar o óbvio, é melhor se preparar para isso do que lidar com isso na hora. Um sistema de backup em grande escala precisa fornecer as seguintes capacidades:

* **Alto Desempenho:** O backup precisa se encaixar no prazo alocado. Isso pode incluir backups online que não impactam significativamente o desempenho, incluindo partições em tempo real.
* **Administração Simples:** A interface de administração deve fornecer ferramentas que permitam identificar facilmente objetos a serem copiados (incluindo tabelas, espaços de tabelas e logs de redo), criar cronogramas e manter a verificação e os logs de backup para restauração subsequente.
* **Operações Automatizadas:** A facilidade de backup deve fornecer serviços de gerenciamento de armazenamento, agendamento automatizado, manuseio de mídia e dispositivos, relatórios e notificação.

O backup do warehouse é geralmente um backup físico. Esta é uma imagem do banco de dados em um determinado ponto no tempo, incluindo índices e informações de layout físico.

**Arquivamento e Recuperação**

Decidir o que mover para fora do warehouse é uma questão de custo-benefício. Custos de manter os dados por perto—ocupam espaço em disco e desaceleram os tempos de carga e consulta. Por outro lado, os usuários de negócios podem precisar desses dados para fazer algumas análises históricas críticas. Da mesma forma, um auditor pode solicitar dados arquivados como parte de um procedimento de conformidade. A solução não é jogar os dados fora, mas colocá-los em algum lugar que custa menos, mas ainda é acessível. Arquivamento é a segurança do warehouse.

No momento desta escrita, o custo do armazenamento em disco online está caindo tão rapidamente que faz sentido planejar muitas tarefas de arquivamento para simplesmente escrever no disco. Especialmente se o armazenamento em disco for tratado por um recurso de TI separado, a necessidade de "migrar e atualizar" é substituída por "atualizar". Você precisa garantir que pode interpretar os dados em vários pontos no futuro.

Quanto tempo leva para os dados ficarem obsoletos depende da indústria, do negócio e dos dados específicos em questão. Em alguns casos, é bastante óbvio quando os dados mais antigos têm pouco valor. Por exemplo, em uma indústria com evolução rápida de novos produtos e concorrentes, a história não necessariamente ajuda a entender hoje ou prever amanhã.

Depois de determinar arquivar certos dados, a questão se torna "quais são as implicações de longo prazo de arquivar dados?" Obviamente, você precisa aproveitar mecanismos existentes para mover fisicamente os dados de sua mídia atual para outra mídia e garantir que possam ser recuperados, junto com uma trilha de auditoria que contabilize os acessos e alterações nos dados. Mas o que significa "manter" dados antigos? Dadas as crescentes preocupações de auditoria e conformidade, você pode enfrentar requisitos de arquivamento para preservar esses dados por cinco, 10 ou talvez até 50 anos. Que mídia você deve utilizar? Você será capaz de ler essa mídia nos próximos anos? Em última análise, você pode se encontrar implementando um sistema de biblioteca capaz de arquivar e atualizar regularmente os dados, e depois migrá-los para estruturas e mídias mais atuais.

Finalmente, se você estiver arquivando dados de um sistema que não será mais usado, pode ser necessário "desativar" os dados extraindo-os do sistema e escrevendo-os em um formato padrão que seja independente do aplicativo original. Pode ser necessário fazer isso se a licença para usar o aplicativo terminar.

### Subsistema 24: Sistema de Recuperação e Reinício

Depois que o sistema ETL está em produção, falhas podem ocorrer por inúmeras razões além do controle do processo ETL. Causas comuns de falhas de produção ETL incluem:

* Falha de rede
* Falha de banco de dados
* Falha de disco
* Falha de memória
* Falha de qualidade de dados
* Atualização de sistema não anunciada

Para se proteger dessas falhas, você precisa de um sistema de backup sólido (subsistema 23) e um sistema de recuperação e reinício complementar. Você deve planejar para erros irrecuperáveis durante a carga porque eles acontecerão. O sistema deve antecipar isso e fornecer recuperação de falhas, parada e capacidade de reinício. Primeiro, procure por ferramentas apropriadas e processos de design para minimizar o impacto de uma falha. Por exemplo, um processo de carga deve confirmar conjuntos de registros relativamente pequenos por vez e acompanhar o que foi confirmado. O tamanho do conjunto deve ser ajustável porque o tamanho da transação tem implicações de desempenho em diferentes SGBDs.

O sistema de recuperação e reinício é usado, claro, para retomar uma tarefa que foi interrompida ou para reverter toda a tarefa e reiniciá-la. Este sistema é significativamente dependente das capacidades do sistema de backup. Quando uma falha ocorre, a reação inicial é tentar salvar o que foi processado e reiniciar o processo a partir desse ponto. Isso requer uma ferramenta ETL com funcionalidade de checkpoint sólida e confiável, para que ela possa determinar perfeitamente o que foi processado e o que não foi para reiniciar a tarefa no ponto exato. Em muitos casos, pode ser melhor reverter todas as linhas que foram carregadas como parte do processo e reiniciar desde o início.

Recomendamos frequentemente projetar tabelas de fatos com uma chave surrogada primária de coluna única. Esta chave surrogada é um inteiro simples que é atribuído em sequência à medida que as linhas são criadas para serem adicionadas à tabela de fatos. Com a chave surrogada da tabela de fatos, você pode facilmente retomar uma carga que foi interrompida ou reverter todas as linhas na carga restringindo um intervalo de chaves surrogadas.

**NOTA:** Chaves surrogadas de tabela de fatos têm vários usos na retaguarda ETL. Primeiro, como descrito anteriormente, elas podem ser usadas como base para reverter ou retomar uma carga interrompida. Segundo, fornecem identificação imediata e inequívoca de uma única linha de fatos sem necessidade de restringir múltiplas dimensões para buscar uma linha única. Terceiro, atualizações às linhas de fatos podem ser substituídas por inserções mais exclusões porque a chave surrogada da tabela de fatos é agora a chave real para a tabela de fatos. Assim, uma linha contendo colunas atualizadas pode ser inserida na tabela de fatos sem sobrescrever a linha que deve substituir. Quando todas essas inserções são concluídas, então as linhas antigas subjacentes podem ser excluídas em uma única etapa. Quarto, a chave surrogada da tabela de fatos é uma chave pai ideal para ser usada em um design de pai/filho. A chave surrogada da tabela de fatos aparece como uma chave estrangeira no filho, junto com as chaves estrangeiras das dimensões do pai.

Quanto mais tempo um processo ETL leva, mais você deve estar ciente das vulnerabilidades devido a falhas. Projetar um sistema ETL modular composto de processos eficientes que são resilientes contra falhas e terminais inesperados pode reduzir o risco de uma falha resultando em um esforço de recuperação massivo. Consideração cuidadosa de quando estagiar fisicamente os dados escrevendo-os no disco, junto com pontos de recuperação cuidadosamente elaborados e carimbos de data/hora de carga ou chaves surrogadas sequenciais da tabela de fatos, permite que você especifique a lógica de reinício apropriada.

### Subsistema 25: Sistema de Controle de Versão

O sistema de controle de versão é uma capacidade de "snapshoting" para arquivar e recuperar toda a lógica e metadados do pipeline ETL. Ele controla o processamento de check-out e check-in para todos os módulos e tarefas ETL. Deve suportar comparações de origem para revelar diferenças entre versões. Este sistema fornece uma função de bibliotecário para salvar e restaurar o contexto ETL completo de uma única versão. Em certos ambientes altamente regulamentados, será igualmente importante arquivar o contexto completo do sistema ETL ao lado dos dados arquivados e de backup relevantes. Note que números de versão mestre precisam ser atribuídos para o sistema ETL geral, assim como números de versão de lançamento de software.

**NOTA:** Você tem um número de versão mestre para cada parte do sistema ETL, bem como um para o sistema como um todo, não é? E você pode restaurar o contexto completo de metadados do ETL de ontem se houver um grande erro na versão atual? Obrigado por nos tranquilizar.

### Subsistema 26: Sistema de Migração de Versão

Depois que a equipe ETL supera o processo difícil de projetar e desenvolver o processo ETL e completa a criação das tarefas necessárias para carregar o data warehouse, as tarefas devem ser agrupadas e migradas para o próximo ambiente — do desenvolvimento para o teste e finalmente para a produção — de acordo com o ciclo de vida adotado pela organização. O sistema de migração de versão precisa se integrar ao sistema de controle de versão para controlar o processo e reverter uma migração, se necessário. Ele deve fornecer uma única interface para definir informações de conexão para toda a versão.

A maioria das organizações isola os ambientes de desenvolvimento, teste e produção. Você precisa ser capaz de migrar uma versão completa do pipeline ETL do desenvolvimento para o teste e finalmente para a produção. Idealmente, o sistema de teste é configurado de maneira idêntica ao seu sistema de produção correspondente. Tudo o que é feito no sistema de produção deve ter sido projetado no desenvolvimento e o script de implantação testado no ambiente de teste. Cada operação de retaguarda deve passar por uma rigorosa criação de scripts e testes, seja a implantação de um novo esquema, a adição de uma coluna, a alteração de índices, a alteração do design de agregados, a modificação de um parâmetro de banco de dados, o backup ou a restauração. Operações de front office gerenciadas centralmente, como a implantação de novas ferramentas de BI, a implantação de novos relatórios corporativos e a alteração de planos de segurança, devem ser igualmente rigorosamente testadas e scriptadas se as ferramentas de BI permitirem.

### Subsistema 27: Monitoramento de Workflow

Data warehouses bem-sucedidos estão consistentemente e confiavelmente disponíveis, conforme acordado com a comunidade de negócios. Para alcançar esse objetivo, o sistema ETL deve ser constantemente monitorado para garantir que os processos ETL estejam operando eficientemente e que o warehouse esteja sendo carregado de forma consistente e pontual. O agendador de tarefas (subsistema 22) deve capturar dados de desempenho toda vez que um processo ETL é iniciado. Esses dados fazem parte dos metadados do processo capturados no sistema ETL. O monitor de workflow aproveita os metadados capturados pelo agendador de tarefas para fornecer um painel e um sistema de relatórios que consideram muitos aspectos do sistema ETL. Você desejará monitorar o status das tarefas para todas as execuções de tarefas iniciadas pelo agendador de tarefas, incluindo tarefas pendentes, em execução, concluídas e suspensas, e capturar os dados históricos para suportar a análise de tendências de desempenho ao longo do tempo. Medidas de desempenho chave incluem o número de registros processados, resumos de erros e ações tomadas. A maioria das ferramentas ETL captura as métricas para medir o desempenho do ETL. Certifique-se de disparar alertas sempre que uma tarefa ETL demorar significativamente mais ou menos tempo para ser concluída do que indicado pelo histórico.

Em combinação com o agendador de tarefas, o monitor de workflow também deve rastrear o desempenho e capturar medições do desempenho dos componentes de infraestrutura, incluindo uso de CPU, alocação e contenção de memória, utilização e contenção de disco, uso de pool de buffer, desempenho de banco de dados e utilização e contenção de servidor. Muitas dessas informações são metadados de processo sobre o sistema ETL e devem ser consideradas como parte da estratégia geral de metadados (subsistema 34).

O monitor de workflow tem um papel estratégico mais significativo do que você pode suspeitar. Ele é o ponto de partida para a análise de problemas de desempenho em todo o pipeline ETL. Gargalos de desempenho no ETL podem ocorrer em muitos lugares, e um bom monitor de workflow mostra onde os gargalos estão ocorrendo. O Capítulo 20 discute muitas maneiras de melhorar o desempenho no pipeline ETL, mas esta lista está mais ou menos ordenada começando com os gargalos mais importantes:

* Consultas mal indexadas contra um sistema de origem ou tabela intermediária
* Sintaxe SQL causando escolha errada do otimizador
* Memória de acesso aleatório (RAM) insuficiente causando thrashing
* Ordenação no SGBD
* Etapas de transformação lentas
* I/O excessivo
* Escritas desnecessárias seguidas de leituras
* Exclusão e reconstrução de agregados do zero em vez de incrementalmente
* Filtragem (captura de dados alterados) aplicada tarde demais no pipeline
* Oportunidades não exploradas de paralelização e pipeline
* Log de transações desnecessário, especialmente se estiver fazendo atualizações
* Tráfego de rede e sobrecarga de transferência de arquivos

### Subsistema 28: Sistema de Ordenação

Certos processos comuns de ETL exigem que os dados sejam ordenados em uma ordem específica, como agregação e junção de fontes de arquivos planos. Como a ordenação é uma capacidade fundamental de processamento ETL, ela é destacada como um subsistema separado para garantir que receba a devida atenção como um componente da arquitetura ETL. Há uma variedade de tecnologias disponíveis para fornecer capacidades de ordenação. Uma ferramenta ETL pode, sem dúvida, fornecer uma função de ordenação, o SGBD pode fornecer ordenação via a cláusula SQL SORT, e há várias utilidades de ordenação disponíveis.

Ordenar arquivos de texto delimitados simples com um pacote de ordenação dedicado é incrivelmente rápido. Esses pacotes normalmente permitem que uma única operação de leitura produza até oito saídas ordenadas diferentes. A ordenação pode produzir agregados onde cada linha de quebra de uma determinada ordenação é uma linha para a tabela de agregados, e ordenar mais contar é frequentemente uma boa maneira de diagnosticar problemas de qualidade de dados.

A chave é escolher o recurso de ordenação mais eficiente para suportar os requisitos dentro da sua infraestrutura. A resposta fácil para a maioria das organizações é simplesmente utilizar a função de ordenação da ferramenta ETL. No entanto, em algumas situações pode ser mais eficiente usar um pacote de ordenação dedicado; embora os fornecedores de ETL e SGBD aleguem ter compensado grande parte das diferenças de desempenho.

### Subsistema 29: Analisador de Linhagem e Dependência

Dois elementos cada vez mais importantes solicitados ao sistema ETL são a capacidade de rastrear tanto a linhagem quanto as dependências dos dados no sistema DW/BI:

* **Linhagem:** Começando com um elemento de dados específico em uma tabela intermediária ou relatório de BI, identifique a origem desse elemento de dados, outras tabelas intermediárias upstream contendo esse elemento de dados e suas fontes, e todas as transformações que esse elemento de dados e suas fontes passaram.
* **Dependência:** Começando com um elemento de dados específico em uma tabela de origem ou uma tabela intermediária, identifique todas as tabelas intermediárias downstream e relatórios finais de BI contendo esse elemento de dados ou suas derivações e todas as transformações aplicadas a esse elemento de dados e suas derivações.

A análise de linhagem é frequentemente um componente importante em um ambiente altamente regulamentado onde você deve explicar o fluxo de processamento completo que alterou qualquer resultado de dados. Isso significa que o sistema ETL deve exibir as fontes físicas finais e todas as transformações subsequentes de qualquer elemento de dados selecionado, escolhido tanto do meio do pipeline ETL quanto em um relatório final entregue. A análise de dependência é importante ao avaliar mudanças em um sistema de origem e os impactos downstream no data warehouse e no sistema ETL. Isso implica na capacidade de exibir todos os elementos de dados downstream afetados e campos de relatórios finais afetados por uma possível mudança em qualquer elemento de dados selecionado, escolhido tanto no meio do pipeline ETL quanto em uma origem original (dependência).

### Subsistema 30: Sistema de Escalonamento de Problemas

Normalmente, a equipe ETL desenvolve os processos ETL e a equipe de garantia de qualidade os testa completamente antes de serem entregues ao grupo responsável pelas operações diárias dos sistemas. Para que isso funcione, a arquitetura ETL precisa incluir um sistema de escalonamento de problemas proativamente projetado, semelhante ao que está em vigor para outros sistemas de produção.

Depois que os processos ETL foram desenvolvidos e testados, o primeiro nível de suporte operacional para o sistema ETL deve ser um grupo dedicado a monitorar aplicativos de produção. A equipe de desenvolvimento ETL se envolve apenas se a equipe de suporte operacional não puder resolver um problema de produção.

Idealmente, você desenvolveu processos ETL, envolveu-os em um agendador automatizado e tem capacidades robustas de monitoramento de workflow espiando nos processos ETL enquanto eles são executados. A execução do sistema ETL deve ser uma operação autônoma. Deve funcionar como um relógio sem intervenção humana e sem falhas. Se um problema ocorrer, o processo ETL deve notificar automaticamente o sistema de escalonamento de problemas de qualquer situação que necessite de atenção ou resolução.

Esse feed automático pode assumir a forma de logs de erro simples, mensagens de notificação de operador, mensagens de notificação de supervisor e mensagens de desenvolvedor de sistema. O sistema ETL pode notificar um indivíduo ou um grupo, dependendo da gravidade da situação ou dos processos envolvidos. As ferramentas ETL podem suportar uma variedade de capacidades de mensagens, incluindo alertas por e-mail, mensagens de operador e notificações para dispositivos móveis.

Cada evento de notificação deve ser escrito em um banco de dados usado para entender os tipos de problemas que surgem, seu status e resolução. Esses dados fazem parte dos metadados de processo capturados pelo sistema ETL (subsistema 34). Você precisa garantir que procedimentos organizacionais estejam em vigor para escalonamento adequado, para que cada problema seja resolvido apropriadamente.

Em geral, a estrutura de suporte para o sistema ETL deve seguir uma estrutura de suporte padrão. Primeiro, o suporte de nível é tipicamente uma central de atendimento que é o primeiro ponto de contato quando um usuário nota um erro. A central de atendimento é responsável pela resolução sempre que possível. Se a central de atendimento não puder resolver o problema, o suporte de segundo nível é notificado. Este é tipicamente um administrador de sistemas ou DBA na equipe técnica de controle de produção capaz de suportar falhas gerais de infraestrutura. O gerente de ETL é o suporte de terceiro nível e deve ser conhecedor para suportar a maioria dos problemas que surgem no processo de produção ETL. Finalmente, quando tudo mais falha, o desenvolvedor de ETL deve ser chamado para analisar a situação e ajudar na resolução.

### Subsistema 31: Sistema de Paralelização/Pipeline

O objetivo do sistema ETL, além de fornecer dados de alta qualidade, é carregar o data warehouse dentro da janela de processamento alocada. Em grandes organizações com enormes volumes de dados e um grande portfólio de dimensões e fatos, carregar os dados dentro dessas restrições pode ser um desafio. O sistema de paralelização/pipeline fornece capacidades para permitir que o sistema ETL entregue dentro dessas restrições de tempo. O objetivo deste sistema é aproveitar múltiplos processadores ou recursos de computação em grade comumente disponíveis. É altamente desejável e, em muitos casos, necessário que a paralelização e o pipeline sejam automaticamente invocados para cada processo ETL, a menos que condições específicas impeçam seu processamento de tal maneira, como esperar uma condição no meio do processo.

A paralelização é uma técnica poderosa de desempenho em cada estágio do pipeline ETL. Por exemplo, o processo de extração pode ser paralelizado por partição lógica em intervalos de um atributo. Verifique se o SGBD de origem lida com a paralelização corretamente e não gera processos conflitantes. Se possível, escolha uma ferramenta ETL que lide com a paralelização de processos de transformação intermediários automaticamente. Em algumas ferramentas, é necessário criar processos paralelos manualmente. Isso é aceitável até que você adicione processadores adicionais, e o sistema ETL então não pode aproveitar as maiores oportunidades de paralelização, a menos que você modifique os módulos ETL manualmente para aumentar o número de fluxos paralelos.

### Subsistema 32: Sistema de Segurança

A segurança é uma consideração importante para o sistema ETL. Uma violação de segurança séria é muito mais provável de vir de dentro da organização do que de alguém hackeando de fora. Embora não gostemos de pensar nisso, os membros da equipe ETL apresentam tanto uma ameaça potencial quanto qualquer grupo dentro da organização. Recomendamos administrar a segurança baseada em função em todos os dados e metadados no sistema ETL. Para suportar requisitos de conformidade, você pode precisar provar que uma versão de um módulo ETL não foi alterada ou mostrar quem fez alterações em um módulo. Você deve impor acesso autorizado abrangente a todos os dados e metadados ETL por indivíduo e função. Além disso, você desejará manter um registro histórico de todos os acessos a dados e metadados ETL por indivíduo e função. Outra questão a ser cautelosa é o processo de movimentação de dados em massa. Se você mover dados pela rede, mesmo que seja dentro do firewall da empresa, vale a pena ser cauteloso. Certifique-se de usar criptografia de dados ou uma utilidade de transferência de arquivos que use um protocolo de transferência seguro.

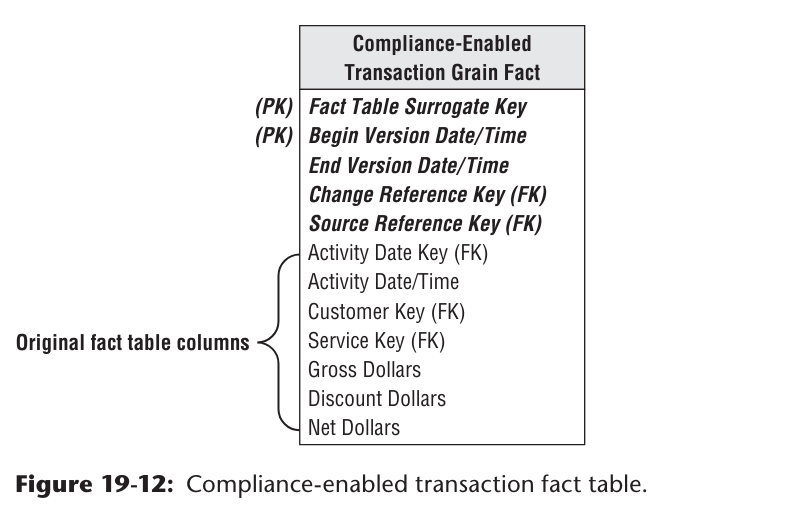
Outra questão de segurança de retaguarda a considerar é o acesso do administrador ao servidor de warehouse de produção e ao software. Vimos situações em que ninguém na equipe tinha privilégios de segurança; em outros casos, todos tinham acesso a tudo. Obviamente, muitos membros da equipe devem ter acesso privilegiado ao ambiente de desenvolvimento, mas o warehouse de produção deve ser estritamente controlado. Por outro lado, alguém da equipe DW/BI precisa ser capaz de redefinir a máquina do warehouse se algo der errado. Finalmente, a mídia de backup deve ser guardada. A mídia de backup deve ter tanta segurança ao seu redor quanto os sistemas online.

### Subsistema 33: Gerente de Conformidade

Em ambientes altamente regulamentados, suportar requisitos de conformidade é um requisito significativo para a equipe ETL. Conformidade no data warehouse envolve "manter a cadeia de custódia" dos dados. Da mesma forma que um departamento de polícia deve manter cuidadosamente a cadeia de custódia de evidências para argumentar que as evidências não foram alteradas ou adulteradas, o data warehouse também deve proteger cuidadosamente os dados sensíveis à conformidade confiados a ele desde o momento em que chegam. Além disso, o data warehouse deve sempre mostrar a condição e o conteúdo exatos desses dados em qualquer ponto no tempo que possa ter estado sob o controle do data warehouse. O data warehouse também deve rastrear quem teve acesso autorizado aos dados. Finalmente, quando o auditor suspeito olhar por cima do seu ombro, você precisa voltar a uma versão arquivada e carimbada com data e hora dos dados como foram originalmente recebidos, que você armazenou remotamente com um terceiro de confiança. Se o data warehouse estiver preparado para atender a todos esses requisitos de conformidade, então o estresse de ser auditado por uma agência governamental hostil ou advogado armado com uma intimação deve ser muito reduzido.

Os requisitos de conformidade podem significar que você não pode realmente alterar nenhum dado, por qualquer motivo. Se os dados precisarem ser alterados, então uma nova versão dos registros alterados deve ser inserida no banco de dados. Portanto, cada linha em cada tabela deve ter carimbos de data/hora de início e término que representem com precisão o período de tempo em que o registro foi a "verdade atual". O grande impacto desses requisitos de conformidade no data warehouse pode ser expresso em termos simples de modelagem dimensional. Mudanças de tipo 1 e tipo 3 estão mortas. Em outras palavras, todas as mudanças se tornam inserções. Não mais exclusões ou sobrescritas.

A Figura 19-12 mostra como uma tabela de fatos pode ser aumentada para que mudanças de sobrescrita sejam convertidas em um equivalente de mudança de tipo 2 na tabela de fatos. A tabela de fatos original consistia nas sete colunas inferiores, começando com a data da atividade e terminando com os dólares líquidos. A tabela de fatos original permitia sobrescritas. Por exemplo, talvez haja uma regra de negócios que atualiza os valores de desconto e dólares líquidos após a criação original da linha. Na versão original da tabela, a história é perdida quando a mudança de sobrescrita ocorre, e a cadeia de custódia é quebrada.



Para converter a tabela de fatos para ser habilitada para conformidade, cinco colunas são adicionadas, conforme mostrado em negrito. Uma chave surrogada da tabela de fatos é criada para cada linha de tabela de fatos não modificada original. Essa chave surrogada, como uma chave surrogada de tabela de dimensão, é apenas um inteiro único que é atribuído à medida que cada linha de tabela de fatos original é criada. O carimbo de data/hora da versão inicial é o momento exato da criação da linha da tabela de fatos. Inicialmente, o carimbo de data/hora da versão final é definido para uma data/hora fictícia no futuro. A referência de mudança é definida como "original" e a referência de origem é definida como a fonte operacional.

Quando uma mudança de sobrescrita é necessária, uma nova linha é adicionada à tabela de fatos com a mesma chave surrogada da tabela de fatos, e as colunas regulares apropriadas são alteradas, como dólares de desconto e dólares líquidos. A coluna de carimbo de data/hora da versão inicial é definida para a data/hora exata quando a mudança no banco de dados ocorre. O carimbo de data/hora da versão final é definido para uma data/hora fictícia no futuro. O carimbo de data/hora da versão final da linha de fato original agora é definido para a data/hora exata quando a mudança no banco de dados ocorre. A referência de mudança agora fornece uma explicação para a mudança, e a referência de origem fornece a fonte das colunas revisadas.

Referindo-se ao design na Figura 19-12, um momento específico no tempo pode ser selecionado e a tabela de fatos restringida para mostrar exatamente o que as linhas continham naquele momento. As alterações em uma determinada linha podem ser examinadas restringindo-se a uma chave surrogada específica da tabela de fatos e classificando por data/hora da versão inicial.

A máquina de conformidade é um acréscimo significativo a uma tabela de fatos normal (consulte a Figura 19-12). Se a tabela habilitada para conformidade for usada apenas para demonstrar conformidade, então uma versão normal da tabela de fatos com apenas as colunas originais pode permanecer como a tabela operacional principal, com a tabela habilitada para conformidade existindo apenas em segundo plano. A tabela habilitada para conformidade não precisa ser indexada para desempenho porque não será usada em um ambiente BI convencional.

Pelo amor de Deus, não assuma que todos os dados agora estão sujeitos a restrições draconianas de conformidade. É essencial que você receba diretrizes firmes do diretor de conformidade antes de tomar qualquer medida drástica.

A base de um sistema de conformidade é a interação de vários dos subsistemas já descritos, casados com algumas tecnologias e capacidades chave:

* **Análise de linhagem:** Mostre de onde veio uma peça final de dados para provar os dados de origem originais mais as transformações, incluindo procedimentos armazenados e mudanças manuais. Isso requer documentação completa de todas as transformações e a capacidade técnica de executar novamente as transformações contra os dados originais.
* **Análise de dependência:** Mostre onde um elemento de dados de origem foi usado.
* **Controle de versão:** Pode ser necessário executar novamente os dados de origem pelo sistema ETL em vigor na época, exigindo a versão exata do sistema ETL para qualquer fonte de dados.
* **Backup e restauração:** Claro, os dados solicitados podem ter sido arquivados anos atrás e precisam ser restaurados para fins de auditoria. Esperamos que você tenha arquivado a versão adequada do sistema ETL junto com os dados, para que tanto os dados quanto o sistema possam ser restaurados. Pode ser necessário provar que os dados arquivados não foram alterados. Durante o processo de arquivamento, os dados podem ser codificados com hash e o hash e os dados separados. Tenha os códigos de hash arquivados separadamente por um terceiro de confiança. Então, quando solicitado, restaure os dados originais, codifique-os com hash novamente e compare com os códigos de hash recuperados do terceiro de confiança para provar a autenticidade dos dados.
* **Segurança:** Mostre quem acessou ou modificou os dados e as transformações. Esteja preparado para mostrar funções e privilégios para usuários. Garanta que o log de segurança não possa ser alterado usando uma mídia de gravação única.
* **Dimensão de auditoria:** A dimensão de auditoria vincula o contexto dos metadados de execução diretamente com os dados para capturar eventos de qualidade no momento da carga.

### Subsistema 34: Gerente do Repositório de Metadados

O sistema ETL é responsável pelo uso e criação de grande parte dos metadados envolvidos no ambiente DW/BI. Parte da estratégia geral de metadados deve ser capturar especificamente os metadados ETL, incluindo metadados de processo, metadados técnicos e metadados de negócios. Desenvolva uma estratégia equilibrada entre não fazer nada e fazer demais. Certifique-se de que haja tempo nas tarefas de desenvolvimento ETL para capturar e gerenciar metadados. E, finalmente, certifique-se de que alguém na equipe DW/BI seja designado como gerente de metadados e tenha a responsabilidade de criar e implementar a estratégia de metadados.

### Resumo

Neste capítulo, apresentamos os principais blocos de construção do sistema ETL. Como você pode apreciar melhor agora, construir um sistema ETL é incomumente desafiador; o sistema ETL deve abordar uma série de requisitos exigentes. Este capítulo identificou e revisou os 34 subsistemas de ETL e reuniu esses subsistemas em quatro áreas chave que representam o processo ETL: extração, limpeza e conformidade, entrega e gerenciamento. A consideração cuidadosa de todos os elementos da arquitetura ETL é a chave para o sucesso. Você deve entender toda a amplitude dos requisitos e, em seguida, definir uma arquitetura apropriada e eficaz. O ETL é mais do que simplesmente extrair, transformar e carregar; é um conjunto de tarefas complexas e importantes. No próximo capítulo, descreveremos os processos e tarefas para construir o sistema ETL.