

Modelagem Dimensional

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

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

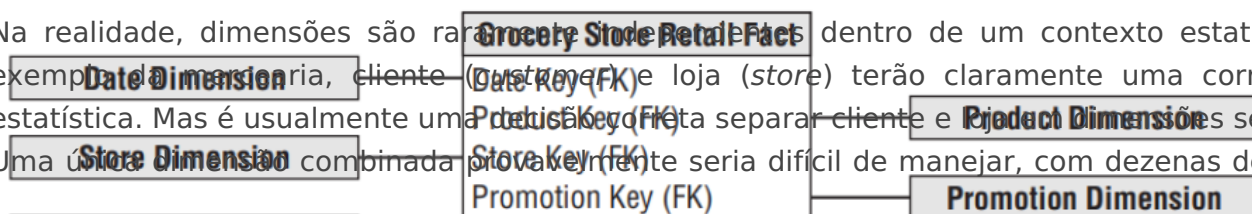
O modelo dimensional deve corresponder à estrutura física dos eventos de captura de dados. Um modelo não deve ser elaborado para atender a um relatório em específico (*report-of-the-day*). Os processos de negócios de uma empresa ultrapassam as fronteiras dos departamentos e funções organizacionais. Em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

Medições e Contexto

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

Embora possamos aglomerar todo o contexto juntamente com cada medida em um vasto e único registro lógico, veremos que, geralmente, é mais conveniente separar o contexto em grupos independentes. Quando armazenados fatos -- vendas em reais de uma mercearia referentes a um produto específico, por exemplo -- naturalmente dividiremos o contexto entre grupos denominados Nome do Produto, Loja, Horário, Cliente, Balconista, dentre outros. Chamamos esses agrupamentos lógicos de *dimensões*, e assumimos informalmente que essas dimensões são independentes. A figura abaixo mostra um modelo dimensional para um armazenamento típico de fatos para uma mercearia (KIMBALL e ROSS, 2010).

Na realidade, dimensões são raras dentro de um contexto estatístico. No exemplo da mercearia, cliente (customer) e loja (store) terão claramente uma correlação estatística. Mas é usualmente uma prática boa e correta separar cliente e produto em dimensões separadas. Uma única dimensão combinada provavelmente seria difícil de manejar, com dezenas de milhões



de linhas. O registro de onde um cliente comprou em determinada loja é expressado mais naturalmente em uma Tabela Fato, que mostra também quando isso ocorreu (dimensão tempo) (KIMBALL e ROSS, 2010).

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

Relacionando os dois Mundos da Modelagem

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

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

Lembre-se que na área de apresentação de um DW não precisamos nos preocupar em forçar a aplicação de regras de dados muito rígidas, exigindo dimensões *snowflaked* (terceira forma normal). A garantia de integridade já é garantida no estágio de ETL (*staging ETL system*) (KIMBALL e ROSS, 2010).

Revisão #11

Criado Fri, Mar 4, 2022 11:20 AM por FLAVIO LOPES DE MORAIS

Atualizado Tue, Mar 8, 2022 8:24 AM por FLAVIO LOPES DE MORAIS