Data Warehouse

De WikiCTBC

ETL > Data Warehouse

Tabela de conteúdo

Data Warehouse

Data Warehouse é um banco cujos dados provindos de diversas fontes passam pelo processo de Extração, Transformação e Carga, tornando estes organizados para melhor análise. O DW é a base dos sistemas de B.I., ele é dividido em quatro elementos bácisos:

Sistemas Operacionais

Todos os Sistemas Transacionais e/ou Sistemas Legados que registram transações na organização, cujos dados serão coletados para o DW.

Staging Area

É um repositório temporário, fora dos limites dos usuários, onde os dados provindos de diversos sistemas serão corrigidos, padronizados e organizados para então serem carregados no DW ou Data Marts. Esse processo é importante, pois como os dados vêm de diversas fontes, podem estar sendo representados com unidades de medidas ou moedas diferentes. Levando em consideração que a Staging Area apenas faz a limpeza, organização e padronização dos dados, ela poderia ser descartada, sendo assim os dados seriam "preparados" em tempo real e enviados diretamenta para o DW. Porém, existem algumas dificuldades para se fazer isso, por exemplo:

- É necessário enviar para o DW dados de duas tabelas em dois bancos de dados fisicamente diferentes. Não é possível fazer uma query que retorne esse resultado, sendo assim esses dados são extraídos para a Staging Area, onde essa query é, então, possível;
- O processo ETL envolve transformações complexas que exigem espaço extra para que esses dados sejam temporariamente armazenados durante essas operações.

Por outro lado:

- Aumenta a latência, que é o tempo necessário para uma mudança nos sistemas de origem ocorra no DW, sendo assim, aplicações em tempo real geralmente evitam o uso da Staging Area;
- E por último, a Staging Area ocupa espaço extra.

Ainda assim, os benefícios gerados pela Staging Area superam seus pontos negativos.

Área de Apresentação

A área de apresentação é composta por Data Marts, que são bases de dados consolidadas orientadas à assuntos específicos como vendas ou movimentação de estoque. Os dados em um Data Mart não são normalizados, o que proporciona melhor performance. O fato de os dados serem desnormalizados, diminui a quantidade de tabelas que, consequentemente, diminui a quantidade de JOINs deixando as pesquisas menos complexas e mais eficientes. A Área de Apresentação é acessada pelos usuários por meio de diversas ferramentas de pesquisa, aplicações analíticas e geradores de relatórios, entre outros.

Ferramentas de Acesso

Podem ser simples ferramentas de pesquisa, geradores de relatórios, aplicações analíticas ou mesmo sofisticadas aplicações para mineiração de dados.

ETL

ETL (Extract, Transform, Load) é o processo de coletar dados de diversas fontes dentro do ambiente organizacional, transformar esses dados de forma a padronizá-los e por fim fazer sua carga em um Data Mart ou Data Warehouse.

O processo de Extração basicamente coleta dados de diferentes sistemas.

O processo de Transformação realiza atividades como:

- Traduzir dados codificados (Ex.: os dados de gênero coletados indicam 1 para masculino e 2 para feminino, mas o DW armazena como M e F);

- Derivar um valor (Ex.: deve-se armazenar o valor total da venda, mas os dados coletados são quantidade vendida e preço unitário, sendo assim multiplica-se os dois para gerar o dado desejado);
- Agrupar dados semelhantes e evitar duplicatas;
- Agregação (Ex.: resumir dados de vendas em vendas totais por loja ou vendas totais por região);
- Gerar chaves substitutas (Surrogate Keys);
- Identificar Slowly Changing Dimensions;

O processo de Carga envia dados para o seu destino, onde serão mantidos de forma centralizada para facilitar o acesso e análise. E como o processo de carga interage diretamente com um banco de dados, as constraints(restrições) nesse banco também vão contribuir para a melhora na qualidade desses dados.

Tabela de Fatos

A tabela de fatos consiste de medidas, métricas ou fatos de um determinado processo de negócio. Ela é localizada no centro de um Star Schema ou Snowflake Schema.

Composição

Nela há dois tipos de colunas, fatos e dimensões. Os fatos pode ser do tipo aditivo, os quais podem ser somados através de qualquer dimensão, como por exemplo as vendas, que são aditivas por produto, período ou loja, semi-aditivo, que podem ser somados através de algumas dimensões, como funcionários que são aditivos por departamento e não por tempo, e não-aditivo, que não podem ser somados através de nenhuma dimensão, como porcentagens.

Há também um tipo de tabela que pode não conter nenhum tipo de medida ou fato, ela é chamada de tabela sem fatos. Ela é composta apenas por chaves estrangeiras para tabelas de dimensão e pode ser usada para registrar eventos, como a frequência de estudantes em um colégio, onde a tabela consistiria do ID do estudante, ID da aula e ID do horário. Aqui você poderia verificar as aulas em que determinado estudante esteve presente, mas não haveria nenhum dado que pudesse ser resumido através dessas dimensões.

As colunas de dimensão constituem a chave primária da tabela de fatos, elas são chaves estrangeiras das tabelas de dimensão definidas por uma surrogate key (chave substituta), o que diminui consideravelmente o tamanho da tabela de fatos, pois dados de texto, as vezes até extensos, são substituidos por um simples valor numérico.

Tipos Fundamentais

Transactional

É a mais básica, sua granularidade é geralmente uma linha por transação, como, por exemplo, uma linha em uma ordem de compra, que consiste no produto, preço unitário e quantidade.

Periodic Snapshot

Registra uma imagem, ou o estado de determinado processo em um momento específico, como saldo bancário ou nível de estoque ao fim do mês.

Accumulating Snapshot

Cada linha representa o ciclo de vida de um processo, por exemplo, uma ordem, desde sua criação até o

processamento completo, ou um processo de admissão. Esse tipo de tabela é composto por diversas dimensões de data, onde cada um delas representa o momento em que determinada fase do processo foi concluída, sendo assim, quando uma fase estiver concluída, a linha desse registro na tabela de fatos deve ser visitada e atualizada.

Tabela de Dimensão

Ao contrátrio das tabelas de fatos, as tabelas de dimensão contem atributos descritivos, geralmente campos de texto, que são usados para filtrar pesquisas e rotular relatórios. Seus valores devem ser verbosos, descritivos, completos, sem códigos ou abreviações. Elas são identificadas por um valor numérico (surrogate key), o que proporciona uma melhor performance nas pesquisas, evita transtornos causados por códigos que eventualmente são reusados por sistemas operacionais e auxilia na integração de dados originados de diferentes fontes.

Uma tabela de dimensão com poucos atributos, embora fácil de dar manutenção, resulta na necessidade de combinar diversas tabelas ao fazer uma pesquisa, deixando as queries mais complexas.

Dimensões Conformadas

O objetivo de uma tabela de dimensão é prover padronização, uma dimensão conformada pode ser compartilhada através do ambiente do Data Warehouse da empresa, possibilitando a junção de diversas tabelas de fatos que representam diferentes processos de negócio. Dimensões conformadas proporcionam:

- Consistência – todas as tabelas de fatos são rotuladas igualmente;
- Integração – é possível pesquisar diferentes tabelas de fatos e então combinar os resultados com atributos de dimensões conformadas;
- Eficiência – dimensões comuns estão sempre disponíveis, sem a necessidade de recriá-las.

Slowly Changing Dimensions

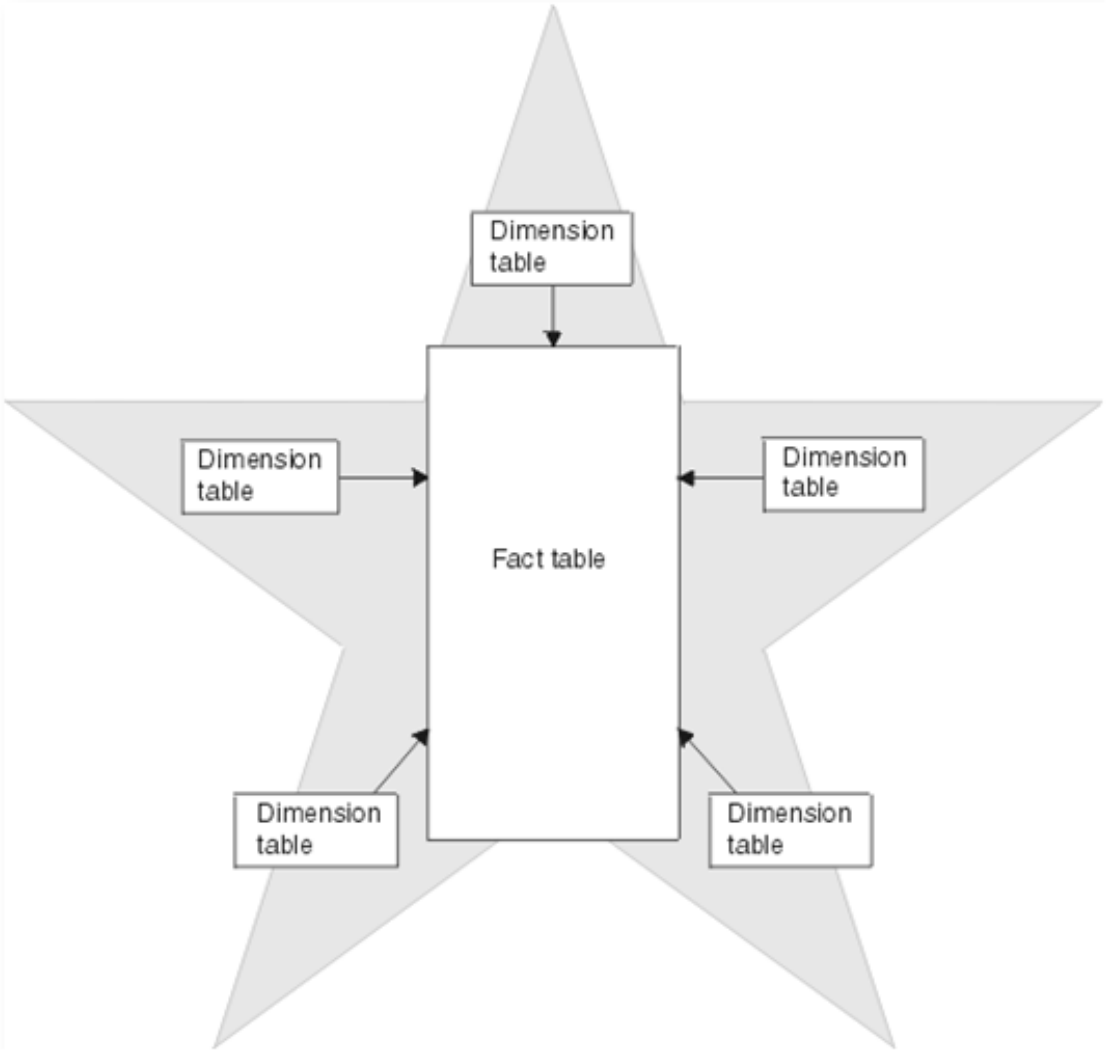
Com o tempo, determinado atributo em uma tabela de dimensão pode mudar, esse fenômeno é chamado de Slowly Changing Dimensions por Kimball. Existem três estratégias para lidar com esse tipo de mudança:

1. Sobrescrever o campo;
2. Adicionar uma nova linha com o novo valor, mantendo as duas versões, porém distinguindo-as com números de versão, ou datas de início e fim, sendo que na última, a data de fim será nula;
3. Criar um novo atributo, que será uma coluna a mais na tabela de fatos, o valor antigo pode ser descrito como valor original, e o novo como valor atual;

Star Schema

Um esquema estrela é composto por uma tabela de fatos em seu centro e diversas tabelas de dimensão ao redor dela.

Esse modelo vai representar um evento, como, por exemplo, uma venda, onde a tabela de fatos, no centro, indica preços e quantidades, enquanto as tabelas de dimensão descrevem a venda, indicando modelo do produto, cor do produto, vendedor responsável, loja onde ocorreu a venda, entre outros.



O Star Schema é desnormalizado, gerando benefícios como:

Queries mais simples

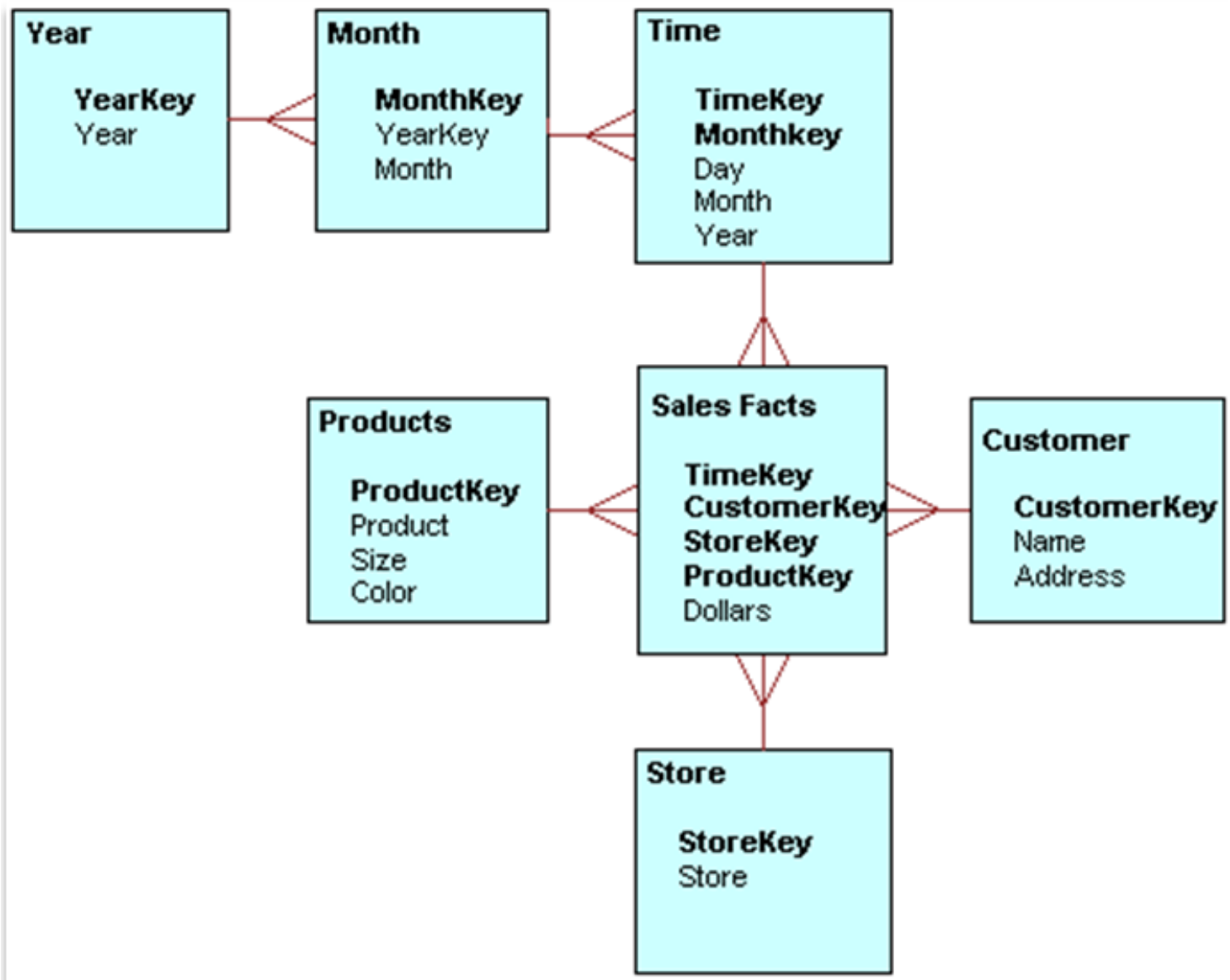
Pelo fato da não-normalização, os número de tabelas é menor, o que proporcia a capacidade de construir queries mais simples para fazer junções.

Agregações mais rápidas

As queries mais simples em um Star Schema resultam em melhor performance nas operações de agregação.

Snowflake Schema

O Snowflake Schema é uma variação do Star Schema, onde as tabelas de dimensão são normalizadas. Essa normalização é realizada removendo-se dados redundantes e formando novas tabelas.
Exemplo:



Uma dimensão de data que consiste em Ano, Mês e Dia, é dividida em três tabelas, onde a tabela Ano é conectada à tabela Mês, que é conectada à tabela Dia, respectivamente.

Vantagens e Desvantagens

Embora esse modelo aumente o número de tabelas deixando as queries mais complexas, ele proporciona mais velocidade de resposta nas pesquisas.

Outros tipos de Dimensão

Degenerate Dimension

É um atributo de dimensão localizado na tabela de fatos que não possui tabela de dimensão própria. São chaves naturais como números de ordens, tickets ou transações de cartões de crédito. Cada transação tem várias linhas, porém, apenas um código que a identifica, e como não há necessidade de criar uma tabela específica para esse valor, ele é adicionado a tabela de fatos, podendo ser usado para agrupar registros de uma mesma transação.

Role-Playing Dimension

Refere-se à uma dimensão que tem diversas funções em uma mesma tabela de fatos. O exemplo mais utilizado é o da dimensão de data, que pode, em uma tabela de fatos de ordem, representar data de criação da ordem, data de embarque, data prevista para entrega, data efetiva da entrega, etc. Sendo assim, a tabela de fatos tem várias chaves estrangeiras referentes à data, porém elas não apontam para a tabela de dimensão, e sim para Views dela, que representam a data de cada etapa do processamento da ordem.

Minidimension

Uma tabela de dimensão que contém atributos que mudam com frequência pode crescer descontrolavelmente se essas mudanças forem realizadas com o Tipo 2 usado na Slowly Changing Dimension. Para evitar que isso aconteça, a dimensão em questão é dividida em minidimensões (as quais são referenciadas por uma Chave Estrangeira na tabela de fatos), onde cada uma dessas minidimensões vai conter os atributos que mudam com frequência.

Outrigger

Uma Outrigger é usada quando uma dimensão é normalizada. Sendo assim, ela contem atributos que são compartilhados por mais de uma dimensão. A diferença em uma Outrigger é uma Minidimension é que a Outrigger é referenciada por uma chave estrangeira nas tabelas de dimensão, e a Minidimension é referenciada por uma chave estrangeira nas tabelas de fato.

Junk Dimension

Uma Junk Dimension pode ser comparada à uma gaveta onde se guarda diversos objetos que, devido a suas quantidades, não há necessidade de fornecer locais próprios para armazená-los, como, por exemplo, clips de papel, fitas adesivas, canetas, entre outros. Sendo assim, a Junk Dimension vai conter atributos como indicadores(sim/não,verdadeiro/falso), comentários opcionais, e outros. Considerando que ela tenha três atributos, a tabela de fatos teria uma diminuição em suas chaves estrangeiras de N para (N – 2), pois as 3 chaves que indicavam dimensões agora agrupadas em uma Junk Dimension, vão se resumir à uma chave apenas. Benefícios gerados são a diminuição do número de dimensões e a diminuição da largura da linha na tabela de fatos.

Referências

1. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling 2ª Edition

- Alerta de Conteúdo