Data Wharehouse

Como o nome sugere, Data Warehouse (DW), em português, é um armazém de dados. Segundo (INMON, 2002) um Data Warehouse é uma coleção de dados de uma coorporação que tem como objetivo dar suporte a tomada de decisão. O DW tem como caracteristica ser:

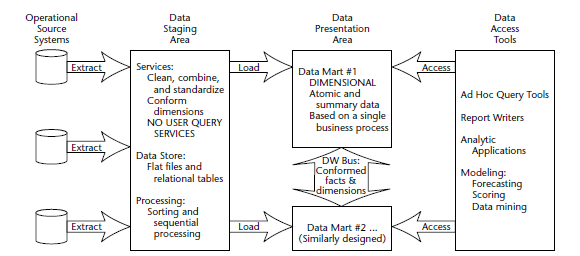
* Integrado: capaz de integrar dados de diversas fontes de formatos.
* Orientado a assunto: um sistema corporativo pode fornecer diversas informações sobre determinados vários aspectos da corporação. O DW é construído focando em alguns aspectos específicos, como por exemplo, em um sistema de vendas de uma loja, podemos analisar as informações sobre venda, como também podemos analisar informações sobre frequência de funcionários ou também sobre o estoque. Cada aspecto sugere a seleção de apenas dados específicos do sistema que serão uteis para a análise desse aspecto. Os demais dados não são de interesse desse aspecto (assunto).
* Não volátil: os dados de um DW representam a informação capturada em um determinado momento da aplicação. Em aplicações, os dados estão sempre sujeitos a modificações. Dessa forma, o DW irá capturar novamente, em outro momento, essa informação e irá acrescentá-la, e não atualiza-la, ao DW, permitindo visualizar a atualização dessa determinada informação em momentos diferentes.
* Temporais: consiste em armazenar a data referente à informação ou à coleta da informação. Dessa forma, o DW consegue fornecer várias visões da informação agrupadas por medidas de tempo.

Data warehousing

O conjunto de ferramentas de manipulação dos dados, desde sua extração até a sua visualização para o apoio a consultas e tomada de decisão é denominado de Data Warehousing. Portanto, DWing não são as tecnologias em si envolvidas e sim uma arquitetura que requer o suporte de diferentes tipos de tecnologias (Inmon, W. H, 2002). As principais tecnologias envolvidas em um ambiente de DWing são:

* SGBDS – Gerenciadores de bases de dados
* Sistemas de conversão e transformação de dados (ferramentas de ETL)
* Tecnologias cliente e servidor para dar acesso aos dados a múltiplos clientes
* Ferramentas de análise e geração de relatórios.

(Kimball, 2002) define os componentes básicos de um ambiente de dwing da seguinte forma:



**Sistemas de fonte de dados Operacionais**

O sistemas de fonte de dados Operacionais (OSS) são as fontes dos dados de negócio que irá compor o ambiente de DW. Pode ser de origem de várias aplicações que compõe o sistema corporativo de uma instituição, podendo ser unificada com uma única ou várias bases de dados. Essas bases de dados não precisam ser necessariamente da mesma tecnologia, podendo ser um banco de dados transacionais, arquivos de texto, planilhas, arquivos xml ou qualquer outra forma de se armazenar e representar informações de negócio (ACAHAR REFERENCIA PARA ISSO). As fontes de dados não podem ser consideradas dentro do escopo do DWing, pois não se tem nenhum controle sobre o conteúdo ou sobre o formato de dados provenientes da fonte (KIMBALL).

**Área de preparação dos dados**

A área de Preparação dos dados (DAS) é o ambiente na qual é feita a extração, transformação e carga dos dados operacionais, processo comumente chamado de ETL. Essa área de preparação, denominada por (KIMBALL), utiliza arquivos simples ou tabelas relacionais temporários, não acessíveis aos usuários, para armazenamento e manipulação das informações durante o processo de ETL. Assim que os dados estiverem prontos é feito a carga na base de dados dimensional, esta, acessível ao usuário.

A etapa de extração dos dados consiste em ler e entender as fontes de dados e extrair apenas os dados necessários para o DW. Os dados extraidos na camada OSS não são integrados, e segundo (Inmom,2002) esses dados, dessa forma, não podem ser utilizados para dar suporte a uma visão corporativa dos dados, que é uma das essências do ambiente de DWing. Dessa forma, uma série de transformações podem ser feitas buscando a limpeza de dados (resolução de conflitos, tratamento de informações não existente, conversão de dados para um formato padronizado), combinação de dados de diversas fontes, remoção de dados duplicados e atribuições de chaves que serão utilizadas no DW (KIMBALL). Ao final da transformação, as informações necessárias são selecionadas e incluídas na base de dados multidimensional, encontrada na área de apresentação que será tratada em seguida.

**Area de apresentação**

A área de apresentação dos dados (DPA) é onde os dados são organizados, armazenados e disponibilizados para consultas á usuários ou ferramentas de geração de relatórios ou análises. Esse ambiente pode ser materializado no DW em si (KIMBALL).

Um dos propósitos de um DW é ter uma navegação intuitiva e de alta performance nos dados (KIMBALL). A visualização das informações de um DW são resultados de agregações, caracterizados pelas consultas OLAP. Dessa forma, a modelagem relacional, normalmente utilizadas em bancos de dados transacionais, não dão suporte esses propósitos, pois a base de dados é normalizada. O processo de normalização foi inicialmente proposto por Byce Codd que consiste em esquematizar as relações de uma base de dados relacional, minimizando redundância de informações e anomalias de inserção, exclusão e alteração, no que diz respeito a manter a consistência de dados caso haja a duplicidade deste. Dessa forma, uma base de dados relacional normalizada oferece mais segurança em transações de acesso e alteração/inclusão/deleção de dados pois possui esquemas menores e consistentes (SISTEMA DE BANCO DE DADOS ERLAD). O problema de esquemas menores é a performance, pois para obter a informação completa é necessário fazer Join entre esses vários esquemas, e se tratando de uma consulta que envolve grande quantidades de dados, como em um DWing, a performance se torna um fator muito importante para a qualidade do ambiente.

Kimball então propõe o uso da modelagem dimensional, que possui as mesmas informações que as bases normalizadas, porém em formato diferente, atendendo a facilidade na navegação e na performance das consultas.

* Modelagem multidimensional

A modelagem dimensional consiste em ter apenas dois tipos de tabelas: fatos e dimensões.

O modelo dimensional proposto por Kimball é chamado de star scheme. Nele, temos a tabela fato no centro e varias tabelas dimensões se relacionando com essa tabela fato.

FOTO DO ESQUEMA ESTRELa

A idéia da modelagem dimensional é representar os tipos de dados de negócio em estruturas de cubo de dados. As células desse cubo contém os valores medidos e os lados definem as dimensões.

As tabelas fatos são as principais em um modelo dimensional. São elas que armazenam as medidas de negócio, que são instâncias da realidade que podem ser mensuradas de maneira quantitativa. As tabelas fatos consomem cerca de 90% do espaço de uma base de dados dimensional (KIMBALL). As tabelas fatos devem possuir duas ou mais Fks para as chaves primarias de diferentes dimensões. Para juntar as informações basta fazer o join entre elas.

As tabelas dimensões contem as descrições textuais de negócio e ajudam na identificação de um componente da respectiva dimensão. Uma Tabela de dimensão deve ser construída de maneira a incluir atributos que podem ser agregados e fornecer ao usuário mais visões sobre as informações. Como por exemplo, uma tabela de produto tem o atributo “nome do produto” e o atributo “categoria”. Dessa forma, o usuário poderá agrupar os produtos por categoria para ter uma diferente visão sobre a informação. Diferente do modelo relacional, o fato da tabela dimensão não ser normalizada implica na melhoria da performance, pois nesse exemplo citado, “categoria” do produto normalmente seria outra tabela, e para a referida análise seria necessário a realização de um Join, que foi substituído apenas por uma clausula group by.

A modelagem dimensional facilita o processamento analítico dos dados (OLAP) para que sejam transformados em informações, aspecto que será tratado a seguir

* OLAP

O processamento analítico On-Line (OLAP – On-line Analytc Processing) é toda atividade de consulta que busca trazer ao usuário uma visão analítica dos dados através de comparações, visões personalizadas, análises históricas, diferentes cenários e entre outras opções (Kimball). Pode-se definir OLAP como sistemas ou ferramentas que realizam consultas ao DW. Tais sistemas permitem aumentar ou diminuir o nível de detalhes da informação através das seguintes operações:

* + Drill down: Consiste em navegar em uma informação de menor nível de detalhe para uma informação de maior nível de detalhe. Por exemplo, uma análise utilizando a dimensão tempo fornece o tempo por ano. Uma operação de Drill Down consistiria e trazer essa mesma informação por mês.
  + Roll up : Consiste em navegar em uma informação de maior nível de detalhe para um menor nível de detalhes. É exatamente o inverso do Drill down.
  + Slice and disse: A operação de slice consiste em fatiar o cubo, que consiste em selecionar um atributo de uma dimensão específica e olhar apenas as informações das outras dimensões sobre esse atributo, eliminando a dimensão fatiada. Por exemplo, em um cubo com informações sobre a venda, fatiar este cubo pelo ano 2012 consiste em eliminar a dimensão tempo e apenas considerar as vendas para o ano de 2012. A Operação dice, como o nome sugere, consiste em fatiar em formato de cubo. Nessa caso, não será eliminado nenhuma dimensão, mas será selecionado alguns subgrupos em duas ou mais dimensões, resultando em um subcubo. Por exemplo, a operação de dice em um cubo de vendas consiste em selecionar as vendas entre o ano de 2010 a 2013, nas localizações “Rio de janeiro”, “São Paulo”, “Distrito Federal” e produtos de categoria “Alimentos” e “Eletronicos”.
  + Pivoting: Também conhecida como rotate, é uma operação que realiza uma rotação nos eixos de um cubo, gerando uma visualização alternativa da informação.[Cavalcanti, tiago]

**Ferramentas de Acesso de dados – Vizualização de dados**

As Ferramentas de acesso a dados são ferramentas que tem a capacidade de realizar consulta aos dados da Área de apresentação. Elas podem variar desde uma simples ferramentas de consulta ad hoc até ferramentas de análises complexas e de mineração de dados(KIMBALL).

Arquitetura de DWing

Referencias

[1] Inmon, W.H. **Building the Data Warehouse 3 Edição**. John Wiley, 2002.

[2] Kimball

.[Cavalcanti, tiago] [http://www.itnerante.com.br/profiles/blogs/artigo-suporte-a-decis-o-02-sobre-as-opera-es-de-o lap](http://www.itnerante.com.br/profiles/blogs/artigo-suporte-a-decis-o-02-sobre-as-opera-es-de-o%09lap)

Wil M.P. van der Aalst . Artigo Process Cubes: Slicing, Dicing, Rolling Up andDrilling Down Event Data for Process Mining