2
Fundamentação Teórica

Neste capítulo, serão apresentadas as definições e os conteúdos teóricos que fundamentarão a sistemática proposta neste trabalho. O item 2.1 descreve as dificuldades do e-governo para compartilhar os dados públicos de maneira adequada aos inúmeros e variados usuários. O item 2.2 aborda os problemas da transparência necessária nos diversos serviços de governo. O item 2.3 apresenta uma das fontes de informações primordiais que o governo disponibiliza: as estatísticas públicas. O item 2.4 apresenta a tecnologia OLAP como a abordagem adequada para a análise das informações disponibilizadas. Finalmente, o item 2.5 apresenta o processo de Data Warehousing como o processo integrador adequado para gerar a base de dados central com qualidade que garanta análises confiáveis e decisões acertadas sobre esses dados.

2.1.
Governo Eletrônico

Há mais de uma década os governos têm tentado compartilhar informações públicas com seus cidadãos, empresas e outros governos com o objetivo de oferecer maior transparência. Porém, colocar as informações do governo on-line e fazer com que essas informações sejam fáceis de encontrar, disponíveis, acessíveis, compreensíveis e utilizáveis representa um grande desafio, com muitos obstáculos a superar, a fim de cumprir a promessa de um governo mais transparente (W3C, 2009).

Políticas obsoletas, limitações orçamentárias e de pessoal, além de uma cultura burocrática lenta são exemplos de algumas exigências e desafios exclusivos enfrentados pelos governos. Outro desafio comum é a dificuldade de se construir um ambiente tecnológico atual e inovador, pois os governos têm sido lentos em se ajustar aos novos paradigmas de abertura, interação e influência (W3C, 2009).

Algumas perguntas que estão diante do e-Governo: Como é possível integrar novas tecnologias nos sistemas existentes? Até que ponto somos capazes de alcançar efetivamente todos os cidadãos, incluindo aqueles que acessam a Web por intermédio de equipamentos móveis, aqueles com deficiências, ou aqueles sem qualquer acesso à Web?

Estes e outros problemas representam desafios para governos que estão considerando a possibilidade de introduzir ou já estão com o processo de governo eletrônico em curso.

No contexto da globalização, tanto o Brasil quanto os demais países do mundo, investem no e-Governo com maior ou menor intensidade. De acordo com a pesquisa da ONU (UN-Survey, 2010), os países que reconhecem o e-Governo como uma ferramenta poderosa para o desenvolvimento humano consolidam a visão de uma sociedade de informação global. Em contrapartida, os países lentos para abraçar e-governo tendem a ser vistos como países que oferecem serviços baseados na oferta e na procura, países com afastamento entre governo e cidadão, e com processos opacos de tomada de decisões.

O desenvolvimento da Web, a explosão de novas tecnologias e suas práticas associadas oferece aos governos oportunidades para compartilhar às informações públicas. Com o objetivo de ajudar os governos a concretizarem a promessa do e-Governo, a W3C[3] criou o Grupo de Interesse em e-Governo - GI para eGov (W3C, 2009). Uma das atividades desse grupo é atuar nas necessidades e problemas que os governos enfrentam ao oferecer informações eletrônicas, abrindo oportunidades de descoberta, interação e participação .

[3] W3C-World Wide Web Consortium que visa desenvolver padrões para a criação e a interpretação de conteúdos para a Web.

2.2. Transparência

Segundo (OCDE[4], 1961), transparência é um fator vital para o fortalecimento das relações entre o governo e o cidadão. Tal fator pode ser viabilizado através de informação completa, objetiva, confiável, relevante e de fácil acesso e compreensão.

Algumas iniciativas governamentais, citadas anteriormente, vêm sendo colocadas em prática com o intuito de oferecer maior transparência das informações, porém, ainda não há uma definição exata do que é transparência organizacional, nem normas ou modelos que definam práticas ou procedimentos de como estabelecer a transparência dificultando que as organizações coloquem em prática este conceito (Cappelli, 2009).

Em (Cappelli & Leite, 2008), as características dos conceitos existentes foram organizadas e os graus de transparência da informação foram definidos. Com base nessa organização representada na Figura 2, é possível identificar de forma gradativa se as práticas adotadas por uma organização oferecem menos ou mais transparência.

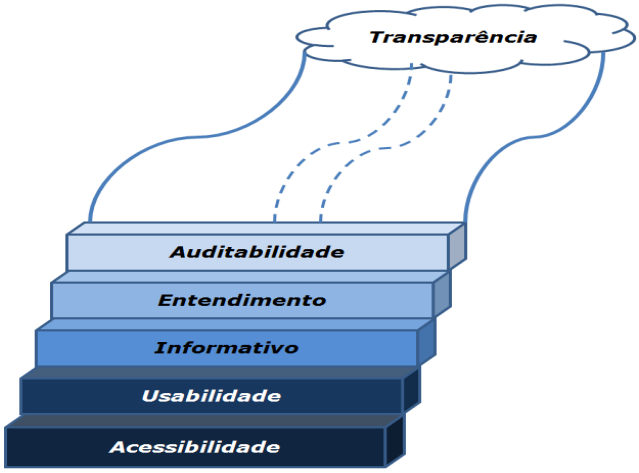

Figura 2 - Degraus da Transparência (Cappelli & Leite, 2008)

[4] OCDE-Organização de Cooperação para o Desenvolvimento Econômico

Cada um dos graus de transparência pode então ser estabelecido através da institucionalização do conjunto de suas características. Assim, pode-se definir os graus como:

Graus	Como a transparência é realizada	Como a capacidade é identificada
GRAU 1 – Acessibilidade	Capacidade de acesso	Práticas que implementam características de portabilidade, operabilidade, disponibilidade, divulgação e desempenho.
GRAU 2 – Usabilidade	Facilidades de uso	Práticas que implementam características de uniformidade, intuitividade, simplicidade, amigabilidade e compreensibilidade.
GRAU 3 – Informativo	Qualidade da informação	Práticas que implementam características de clareza, acurácia, completeza, corretude, consistência e integridade.
GRAU 4 – Entendimento	Entendimento	Práticas que implementam características de composição, concisão, divisibilidade, dependência, adaptabilidade e extensibilidade.
GRAU 5 - Auditabilidade	Auditabilidade	Práticas que implementam características de explicação, rastreabilidade, verificabilidade, validade e controlabilidade

Tabela 1 – Definição de Graus (Cappelli, 2009)

Como pode ser observado, a implementação da transparência é realizada de forma gradativa e pode ser medida através de práticas que implementam as características da informação.

Assim, fazendo uma aferição das práticas adotadas pelos governos para oferecer transparência das informações estatísticas brasileiras, percebe-se que

as estatísticas públicas estão no grau 1. São necessários mecanismos que viabilizem a exploração das estatísticas públicas de forma natural e intuitiva, através de modelos simples, facilitando o uso de forma exploratória das estatísticas públicas , com o objetivo de alcançar o grau 2.

2.3. Estatística Pública

O conjunto das estatísticas de um país, ou de uma região, fornece-nos não apenas a simples constatação e quantificação de determinados fenômenos, mas um conjunto de dados que nos auxiliam a perceber os diversos aspectos que afetam o desenvolvimento de qualquer sociedade (Senra, 1999).

Segundo (Schwartzman, 1996), estatística pública, ou estatística oficial, refere-se à informação estatística produzida pelas agências estatísticas do governo: órgãos de recenseamento, departamentos de estatística e instituições semelhantes. Essas instituições públicas são, simultaneamente, centros de pesquisa, envolvendo, portanto, valores científicos e tecnológicos, além de instituições públicas ou oficiais sujeitas às regras, valores e restrições do serviço público.

A produção sistemática das estatísticas públicas, pelas instituições públicas, é apoiada pela metodologia estatística produzindo um rol de atividades mais ou menos comuns a vários países. Essa convergência tem sido ainda mais acentuada com os avanços tecnológicos e a globalização, que facilita o uso intensificando a necessidade de dispor-se de estatísticas comparáveis entre as diversas nações.

Várias são as ferramentas utilizadas pelas instituições públicas para apoiarem as diferentes etapas de produção de uma estatística pública. Em geral, elas ajudam no planejamento (elaboração, edição do questionário e outras), na entrada dos dados (digitação, importação e outras) e na análise exploratória dos dados com múltiplas possibilidades de tabulações, cruzamentos, testes estatísticos simples ou complexos.

As estatísticas públicas são produzidas para responder às perguntas que os governantes têm sobre as condições de vida da população. A escolha de temas a serem investigados, os conceitos e muitos outros aspectos técnicos e metodológicos para a produção de uma estatística pública é definido pelo "observador-produtor". Assim, ainda que a análise exploratória das estatísticas públicas seja realizada por outras instituições, ela sempre deve ser realizada com base no modelo conceitual definido pela instituição que a produziu.

Conforme afirma Senra (2001), as estatísticas são o resultado de registros coletados individualmente com a intenção de retratar aspectos predeterminados da realidade e, portanto, têm seu sentido e significado, estabelecidos antes de sua coleta.

Após a coleta de dados os institutos de pesquisa geram um conjunto de arquivos sistematizados em diversos formatos denominado de microdados[5], e disponibiliza em seu site para os usuários especialistas ou não. Os microdados contém os dados brutos coletados nas pesquisas, além de arquivos em formatos distintos contendo informações tais como: notas metodológicas da coleta; os objetivos da pesquisa estatística; os metadados para a leitura dos dados brutos, e enfim, a cópia do questionário aplicado as pessoas.

Após a coleta de dados os institutos de pesquisa geram um conjunto de arquivos e dados sistematizados em diversos formatos, que são os microdados, e disponibiliza em seu site. Os microdados contém os dados e as informações necessárias para entendimento dos conceitos aplicados na pesquisa.

Porém, quando uma instituição (ONG, órgão internacional) elabora uma pesquisa estatística e escolhe como fonte de dados uma estatística pública, a instituição se depara com a situação em que os dados, além de estarem distribuídos em diversos microdados, também não estão em formato adequado para análise e interpretação, tornando necessário aplicar certos tratamentos,

[5] Microdados - Segundo o IBGE, Microdados consistem no menor nível de desagregação dos dados de uma pesquisa, retratando, na forma de códigos numéricos, o conteúdo dos questionários, preservado o sigilo das informações. Os microdados possibilitam aos usuários, com conhecimento de linguagens de programação ou softwares de cálculo, criar suas próprias tabelas de dados numéricos. (http://www.ibge.gov.br/censo/divulgacao_digital.shtm acessado em: 03.02.2010)

transformações dos dados e recodificação de variáveis, demandando muito tempo e custo para a realização da pesquisa.

Essa tarefa de integração e preparação dos dados para análise gera informações mais apropriadas e adaptadas à análise, as quais ficam então à disposição do analista para serem consultadas, cruzadas, correlacionadas em relação ao tempo, enfim exploradas. Percebe-se, então, que entre o acesso às estatísticas públicas (microdados) e o seu uso efetivo para análise exploratória para outras pesquisas há um "gap", que é onde esse trabalho se aplica.

Depois que as estatísticas públicas são preparadas para análise, inúmeras são as tecnologias que possibilitam de forma fácil e flexível a exploração das informações. A tecnologia principal de consumo dessas informações é o Processamento Analítico On-Line (OLAP) que é apresentada na seção seguinte.

2.4. OLAP

A característica principal do OLAP é permitir uma visão conceitual multidimensional dos dados. A visão multidimensional é mais natural, fácil e intuitiva, permitindo uma visão dos negócios da organização em diferentes perspectivas (dimensões) tornando assim, o usuário em um explorador de informações (Shoshani, 1997; Campos & Rocha,1997).

As ferramentas OLAP (i.e. ferramentas orientadas para OLAP) são projetadas para apoiar análises e consultas *ad hoc*, além de ajudarem analistas e executivos a sintetizarem informações sobre a organização, através de comparações, visões personalizadas, análise histórica e projeção de dados em vários cenários. Ferramentas OLAP tornaram-se populares em diversas áreas, sendo implementadas para ambientes multi-usuário, arquitetura cliente-servidor e oferecem respostas rápidas e consistentes às consultas interativas executadas pelos analistas, independente do tamanho e complexidade dos dados. (Codd, 1993; Chaudhuri & Dayal, 1997; Inmon, 1999).

O modelo de dados utilizado nas ferramentas OLAP organiza conceitualmente a informação em cubos com diversas dimensões. Esse modelo multidimensional permite a realização de consultas visuais além de considerar a semântica do negócio (Shoshani, 2000).

Cada dimensão do cubo consiste em um conjunto de descritores categóricos organizados em estruturas hierárquicas (Messoud et al., 2004). O usuário pode realizar operações no cubo, agregando dados em dimensões superiores (*roll-up*), desagregando-os, descendo nas inferiores (*drill-down*), ou selecionando e projetando dados (*slice-and-dice*). A abordagem dimensional permite o uso automático de funções de agregação e de consulta visual, além de boa performance e do fato de ser mais natural para a análise de dados.

As atuais tecnologias da Web, baseadas no uso de XML[6] e Serviços Web[7] revolucionaram a troca de dados na Internet possibilitando o desenvolvimento de sites com maior interação. Estas tecnologias têm uma grande abrangência para o intercambio de dados transacionais. Entretanto, os sistemas fundamentados na tecnologia OLAP, onde a estrutura de dados é representada em cubos, declarado por objetos com: dimensões, membros e fatos contrastam com dados transacionais, geralmente representados por tabelas contendo linhas e colunas. Assim, para que aplicações na Web se comuniquem através dos dados estruturados em cubos, a Microsoft e a Hyperion propuseram um protocolo padrão denominado XML for Analysis (XMLA).

O XMLA é baseado em padrões existentes na indústria como o XML, o Serviço Web e o HTTP. A primeira versão das especificações do XMLA foi aprovada em 2001 e atualmente é mantida pelo XMLA *Council*[8], ao qual pertencem agora mais de 25 empresas.

Além do acesso aos dados, o XMLA também oferece acesso aos metadados e esse serviço é especialmente relevante para os propósitos desse

[6] XML- eXtensible Markup Language, linguagem utilizada para descrever dados estruturados.

[7] Serviço Web é, resumidamente, um serviço (ou aplicação independente) que está disponivel na Internet e que usa o padrão XML para troca de mensagens

[8] XMLA *Council* foi criado em abril de 2001 com o objetivo de desenvolver e definir as especificações para facilitar o desenvolvimento de sofisticadas soluções BI

trabalho. Pois, a exploração das estatísticas públicas deve ser realizada sob a semântica de quem as produziu. Com a possibilidade de consultar os metadados semânticos, dos cubos os usuários finais podem realizar suas navegações nas visões de dados no nível dos conceitos de negócio.

O XMLA não é adequado para a troca de cubos de dados completos (dados e metadados) em operações de clientes móveis desconectados, e também não tem suporte nativo para dados geográficos (Dubé et. al, 2009).

O padrão aberto do XMLA descreve dois métodos gerais de acesso: *Discover* e *Execute*. Estes métodos utilizam uma arquitetura cliente-servidor fracamente acoplados que usam o XML para manipular as informações recebidas e enviadas para o servidor Web OLAP.

O método *Discover* obtém dados e metadados de um Serviço Web. Essa informação pode ser a lista de fontes de dados disponíveis ou dados sobre uma fonte específica. O método *Execute* é usado para executar uma consulta MDX ou outro comando específico do servidor XMLA.

A linguagem de consulta do XMLA é a MDX (MultiDimensional eXpression) da mesma forma que a SQL é utilizada para base de dados relacionais. Assim, tanto as consultas MDX, quanto o conjunto de resultados são encapsulados na linguagem XML e são destinados à apresentação nas aplicações cliente-OLAP (ou seja, em planilhas ou aplicações web).

A linguagem MDX permite aos usuários definir novas medidas na consulta em tempo de execução favorecendo a criação de novos indicadores, além dos indicadores previamente definidos no esquema do cubo em análise. Um exemplo de medida calculada que poderia ser gerada no protótipo é a taxa de analfabetismo de uma população.

Através do XMLA, aplicações clientes e servidoras podem ser implementadas em diferentes plataformas de hardware, sistemas operacionais e linguagens de programação. O uso do XMLA garante a interoperabilidade dos componentes OLAP, com a possibilidade de substituição dos componentes por outros de fabricantes diferentes. Surgiu assim o Web Olap ou WOLAP (Arriaga & Marques, 2009).

A utilização de ferramentas WOLAP permite que o acesso aos dados e o seu processamento analítico possa ser feito através de qualquer computador com acesso de rede ao servidor OLAP. O usuário pode acessar os dados através de qualquer ferramenta OLAP que tenha suporte ao XMLA. As ferramentas variam desde aplicações Web, sem a necessidade de instalar algum software específico na máquina do usuário - gratuitas ou pagas-, a ferramentas desktop proprietárias. Além da flexibilidade geográfica proporcionada por esta arquitetura, o custo de manutenção tende a ser inferior a um sistema tradicional de aplicação servidor, uma vez que só é necessário configurar e atualizar os cubos no servidor que fornece este serviço.

Alguns exemplos de ferramentas WOLAP, com suporte a XMLA e "open source" disponíveis são: OpenI, Palo Web Client, SpagoBI e Pentaho BI Suite. Destas, as ferramentas OpenI e Palo Web Client são apenas peças de *software* que permitem a análise dos dados através de um *browser* Web, enquanto as restantes ferramentas SpagoBI e Pentaho BI Suite são suites completas de *Business Intelligence*. Estas suites englobam desde módulos de geração de relatórios a ferramentas de Mineração de Dados[9]. A maioria das ferramentas gratuitas disponíveis na web utilizam JPivot, que é uma biblioteca escrita em java para acesso a bases OLAP (Arriaga & Marques, 2009).

Diante das vantagens e benefícios do uso da tecnologia OLAP para a análise exploratória de dados, inclusive dos avanços na Web citados, torna-se motivador aplicar essa tecnologia para oferecer a usabilidade da informação estatística aos analistas, sociólogos, cidadãos interessados, ONG's e outros.

Porém, para fazer uma análise exploratória em dados que estão em diversas estatísticas públicas é necessário que haja um processo integrador dessas fontes. Neste contexto, esta forma integradora é melhor obtida pela tecnologia de data warehousing (descrita a seguir) que já é bem estabelecida no mercado. Além de favorecer o uso da tecnologia OLAP, data warehousing também favorece o uso de outras aplicações de BI, tais como: geração de relatórios, construção de indicadores, mineração de dados.

[9] Ferramentas de Mineração de Dados são necessárias quando se deseja extrair conhecimento de dados.

2.5. Data Warehousing

O processo de construção, acesso e manutenção de um data warehouse (DW) é denominado de Data Warehousing (DWing). Esse processo objetiva integrar e gerenciar dados copiados de diversas fontes, com o propósito de proporcionar uma visão única de um negócio.

A arquitetura do processo de Ding é voltada para a extração de informação a partir dos dados operacionais de uma organização e exibição desses dados utilizando ferramentas de visualização multidimensionais. A seguir será descrita uma arquitetura ideal para o fluxo de dados ocorra.

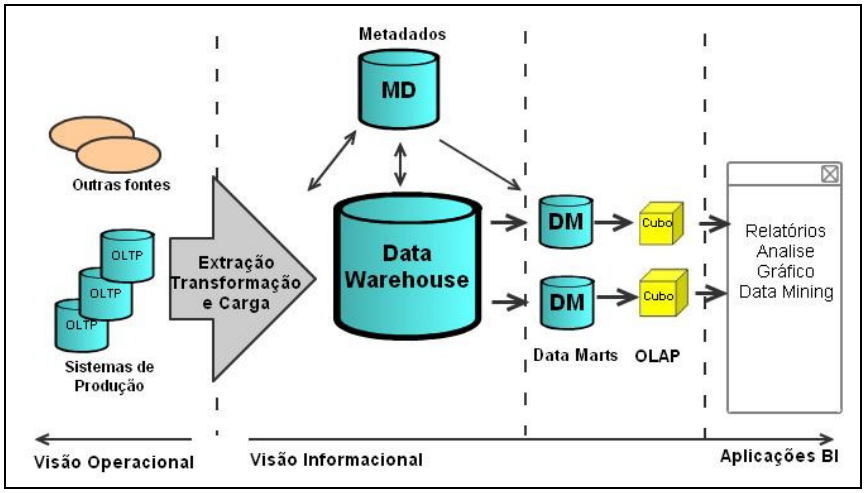

Figura 3 – Arquitetura de DWing (Fonte: autor)

A arquitetura do DWing tem três componentes fundamentais : oETL (*Extract, Transform and Load*), o DW e as aplicações de BI ilustrado na Figura 3.

O processo de Extração, Transformação e Carga (ETL) é considerado crítico para o DW (Kimball, 2008), corresponde à captura dos dados nas diversas fontes, as transformações que agregam valor aos dados, a limpeza e a padronização das diferentes fontes em uma única visão desses dados conforme regras de negócios e por fim a carga dos dados em um DW.

O ETL é um processo meticuloso que tem a importante missão de diagnosticar problemas nos dados, evitando que dados ruins cheguem aos usuários finais. Segundo Oslon (2003), a qualidade dos dados pode ser

garantida com o uso de cenários de qualidade, em que para cada cenário é definido um conjunto de teste a serem executados. Um dos desafios da implantação de um DW é a integração destes dados, eliminando as redundâncias e padronizando as informações para que possam estar representadas no mesmo formato (Berson e Smith, 1997).

Um DW é um banco de dados cuja função é proporcionar aos seus usuários uma única fonte de informação a respeito dos seus negócios, servindo também como ferramenta de apoio ao processo de extração de conhecimento. Além disso, é responsável pelo agrupamento dos dados históricos de uma organização, sejam eles provenientes de qualquer tipo de banco de dados, planilhas eletrônicas, documentos textuais, entre outros. Assim, um DW é um grande repositório de dados, obtidos a partir de várias fontes, que tem diferenças fundamentais em relação aos bancos de dados convencionais (Inmon, 1997).

Todos os componentes de um DW devem ser administrados a partir de um repositório de metadados, auxiliando o administrador e o projetista do DW (Quix, 2000). Para (Kimball, 2008), a integração de todos os metadados em um único repositório é considerada uma solução ideal. Em relação ao conteúdo dos metadados, os administradores do DW, bem como outros usuários de nível técnico, estão interessados, principalmente, nos metadados relacionados à implementação técnica, chamados de metadados técnicos. Já os usuários finais ou de negócios, estão interessados em entender a semântica dos negócios modelados no DW e, por esta razão, necessitam de metadados relacionados à semântica do negócio para formar uma visão orientada aos negócio. (STÖHR *et. al*, 1999)

Após os dados estarem armazenados no DW, eles podem ser transferidos para os *DataMart*(DM)[10](Kimball, 1996) ou estruturas de consulta de alto desempenho voltadas para o atendimento de um grupo específico de usuários. Os DM podem ser acessados, pelos usuários, utilizando as ferramentas OLAP, para buscarem as informações, que são exibidas num formato multidimensional, a qual darão um amplo apoio ao processo de tomada de decisão.

[9] Um data mart é um ambiente analítico voltado a alguma área de negócio específica.

Nessa arquitetura, o DW é uma camada onde os dados provenientes de diversas fontes são integrados, logo, os dados estão num formato mais normalizado. Já o DM é uma estrutura de consulta de alto desempenho, pois os dados estão representados em uma forma desnormalizada, para que as consultas sejam executadas mais eficientemente, pois são necessários menos junções de tabelas quando os dados forem recuperados.

Apesar das consultas serem realizadas sobre o DM, o fato de se ter uma camada de integração evita a repetição da extração, pois é provável que vários DM exijam dados das mesmas fontes. Se os dados não forem trazidos dessas fontes, através de um repositório comum, então cada DM teria que acessar cada fonte. Além do mais, um DW garante uma interpretação padronizada dos dados e fornece um repositório que é bem mais flexível do que as estruturas desnormalizadas dos DM.

Existem diversas alternativas para o DWing como por exemplo, construir somente o DW sem DMs, ou construir apenas os DM sem o DW, ou ainda, não construir nem o DW nem os DMs, e as consultas serem realizadas diretamente nas fontes. Claro que a escolha vai depender das necessidades do projeto.

Existem vários tipos de ferramentas utilizadas para a construção de um DW: ferramentas para armazenamento, extração, transformação e limpeza de dados; repositórios de metadados; transferência de dados e replicação; gerenciamento e administração; e gerenciamento de consultas e de relatórios.

Além dessas, as ferramentas OLAP e as ferramentas utilizadas no processo de Mineração de Dados, são outros tipos de ferramentas que se beneficiam das características de um DW, pois os resultados obtidos com as mesmas auxiliam efetivamente aos tomadores de decisão.

Para que essa arquitetura disponibilize os dados em formato adequado para analise, o projeto e a implementação do DW e DMs requerem conceitos e técnicas de modelagem diferentes dos usados em bancos de dados tradicionais.

2.5.1.
Modelagem Dimensional

A modelagem dimensional é uma técnica utilizada para conceitualização de modelos de negócio. O esquema estrela definido por Kimball (1997) é o mais popular e consiste numa estrutura onde o centro da estrela é ocupado pela entidade fato (medidas numéricas) e a periferia ocupada pelas dimensões. Esse modelo possui três componentes básicos: Fato, Dimensão e Atributos. O Fato representa um elemento, ou uma transação ou um evento associado ao tema da modelagem. A Dimensão é a característica que se quer analisar em cada fato. Os Atributos descrevem as dimensões e fatos. As medidas são atributos ou variáveis numéricas que representam um fato.

O esquema estrela objetiva a desnormalização dos dados, para se obter um melhor desempenho no ambiente de apoio à tomada de decisão, em relação às estruturas altamente normalizadas das bases de dados operacionais. O segredo para se obter esse desempenho é limitar o número de uniões que terão de ser realizadas e a complexidade de cada união. Esse esquema objetiva também a criação de um modelo de dados que seja mais compreensível ao usuário final, procurando representar a maneira natural de como ele enxerga o seu negócio, uma vez que, os esquemas E/R (Entidades/Relacionamentos) são de difícil interpretação por parte dos usuários finais, além de não representar a maneira natural de como eles visualizam seu negócio (Kimball, 1997; Todman, 2001). O modelo Entidade Relacionamento visa a representação dos objetos que devem ser mantidos num sistema. O modelo Multidimensional objetiva a representação de um assunto que vai ser analisado por diversas perspectivas ou dimensões.

A Figura 4 mostra um esquema conceitual multidimensional para Vendas e suas quatro Dimensões para análise: (Tempo, Produto, Cliente e Loja).

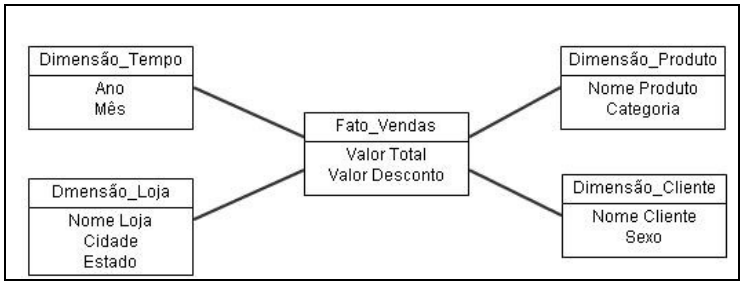

Figura 4 – Exemplo de um esquema multidimensional

No nível lógico de modelagem de um DW, os Fatos, Dimensões e Atributos são representados no Modelo Relacional "estrelado" através de tabelas inter-relacionadas formando esquemas estrela. Em um esquema estrela, uma tabela Fato central interliga por chaves estrangeiras as varias tabelas de Dimensão associadas. Uma variação do esquema estrela, chamado de esquema floco de neve, é utilizado para representar as hierarquias das dimensões através da normalização das tabelas correspondentes. A vantagem desse tipo de esquema é que se torna mais fácil a manutenção das tabelas de dimensão, já que há uma diminuição na redundância dos dados. Entretanto, uma estrutura não normalizada é mais eficiente no momento de execução das consultas e normalmente elas são altamente desnormalizadas, um requisito indispensável em sistemas de DWs (Gatziu & Vavouras, 1999;Golfarelli et al., 1998; Kimball,1997).

Em grandes projetos, o DW normalmente contém vários esquemas estrela com algumas dimensões compartilhadas (dimensões conformes). Quando o esquema de um DW é composto por mais de uma estrutura do tipo estrela, este pode ser chamado de esquema constelação (Barquini, 1996).

As tabelas de dimensão são caracterizadas por vários aspectos gerais. Embora, freqüentemente fala-se que o esquema estrela é desnormalizado, na verdade somente as tabelas de dimensão são desnormalizadas. As tabelas de dimensão possuem mais colunas do que as tabelas do banco de dados operacional e geralmente possuem menos registros do que as tabelas de fatos.

A tabela de fatos é a principal tabela do modelo dimensional lógico. Ela é responsável pelo armazenamento de todas as métricas e chaves relativas a um fato ocorrido. Os campos de uma tabela de fatos são mais úteis quando numéricos e aditivos, pois campos descritivos podem ser encontrados nos atributos das dimensões. Em uma tabela de fatos há basicamente dois tipos de campos: chaves estrangeiras, utilizadas para referenciar as dimensões e os campos de fatos, ou métricas, que têm as "medidas" a serem analisadas (valores, quantidades, etc). Quanto à aditividade (Kimball, 2008), as colunas da tabela de fatos devem ser cuidadosamente definidas pelos projetistas e podem ser classificadas em três grupos: aditivas, semi-aditivas e não-aditivas.

A métrica é aditiva quando faz sentido realizar análises de soma nesta métrica ao longo de qualquer dimensão. Um campo de “Valor da Compra" é um exemplo de métrica aditiva, uma vez que pode ser somado em qualquer dimensão de análise.

As métricas semi-aditivas permitem ser somadas ao longo de algumas dimensões apenas, ou seja, há dimensões em que a soma faz sentido e há dimensões em que não faz. Por exemplo, um campo de “Saldo Bancário" pode ser somado por banco, para saber o disponível total do cliente em um determinado dia considerando todos os seus bancos, ou ainda, somar o saldo de todos os clientes em alguma agência bancária especifica, mas não faz sentido somar o saldo de um mesmo cliente ao longo do tempo.

As métricas não-aditivas não podem ser somadas em nenhuma dimensão. Um bom exemplo deste tipo de métrica são valores percentuais, como por exemplo, a % de margem de lucro. Não faz sentido somar a % de margem de lucro em nenhuma dimensão de análise, pois a soma de um valor percentual não é uma informação útil.

As métricas aditivas são as mais adequadas pois, como podem ser somadas livremente, sua utilização em consultas de agregação não tem nenhuma restrição no DW. Quanto ao nível de detalhamento da tabela de fatos, há três níveis de detalhamento fundamentais para estas tabelas (Kimball, 2008), que são descritos a seguir.

O transacional é o nível de maior detalhamento, onde os dados refletem da maneira mais completa possível a verdade no momento em que a transação ocorreu. As tabelas de fatos neste nível costumam ter um grande número de dimensões associadas.

O instantâneo periódico tem a periodicidade definida, são tiradas “fotografias" das métricas de desempenho e de suas dimensões. Estas fotografias são armazenadas de maneira incremental, ou seja, uma vez inserida, a linha não deve sofrer alterações. Este nível tem um número menor de dimensões e um grande potencial para análise de desempenho, oferecendo uma capacidade analítica ainda maior quando combinadas com tabelas de detalhamento transacional.

No instantâneo incremental também são tiradas “fotografias" mas, ao invés de acumulados, os dados são constantemente atualizados no registro do primeiro fato ocorrido. Esta opção de carga é utilizada quando há um processo com fluxo bem definido, com várias etapas, que não pode ser muito complexo e nem apresentar ciclos, pois todos os campos de todas as etapas já estão na tabela desde a sua primeira ocorrência, mas com valores “genéricos" trazidos na primeira vez, que devem ser atualizados de acordo com a evolução do processo no fluxo.

Os resultados da modelagem multidimensional podem ser implementados diretamente utilizando a tecnologia de banco de dados multidimensional ou através do esquema estrela, em um banco de dados relacional. O conceito de Banco de Bados Multidimensional (BDM) é bem mais simples do que o de banco de dados relacional. Ao invés de armazenar informações como registros em tabelas, BDMs armazenam os dados em *arrays* ou matrizes. Existe no mercado uma classe de SGBDs (Sistema Gerenciador de Banco de Dados) que incorporam a tecnologia de banco de dados multidimensional, são os chamados Sistemas Gerenciadores de Banco de Dados Multidimensionais (SGBDMs).

Projetistas de bancos de dados podem e devem separar o conceito de visão multidimensional dos dados, obtida através da modelagem multidimensional, do conceito de armazenar os dados de forma multidimensional. A falta de um modelo de dados multidimensional convencional, tal como o esquema estrela, para bancos de dados relacionais e a falta de um método de acesso padrão, tal como o SQL, acabaram influenciando a utilização da tecnologia de bancos de dados relacional para representar e armazenar dados multidimensionais (Kimball, 1997).

A Figura 5 mostra um exemplo de esquema modelo lógico dimensional estrelado que traz uma tabela Fato central e quatro tabelas de Dimensões: (Tempo, Produto, Cliente e Loja).

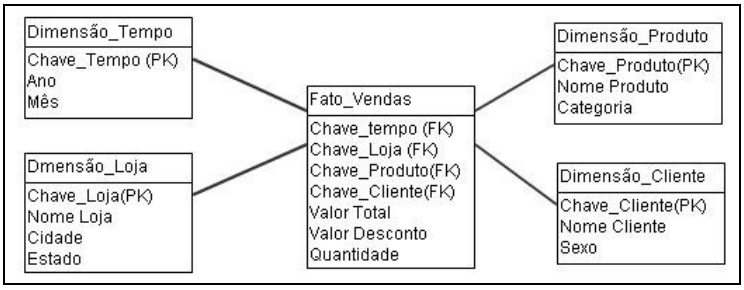

Figura 5 – Exemplo esquema lógico dimensional estrelado

O modelo físico é parecido com o modelo lógico, mas com um maior detalhamento, e mostra a real estrutura que será implementada em algum SGBD. O modelo físico do DW deve ser construído preferencialmente junto com os metadados. A padronização dos nomes dos diversos componentes facilita a manutenção. Fazer uma estimativa de tamanhos também é importante para uma boa definição da estrutura inicial da base.

A Figura 6 mostra o exemplo de um esquema Físico Dimensional estrelado que traz a tabela de Fatos (Vendas) e quatro tabelas de Dimensões: (Tempo, Produto, Cliente e Loja).

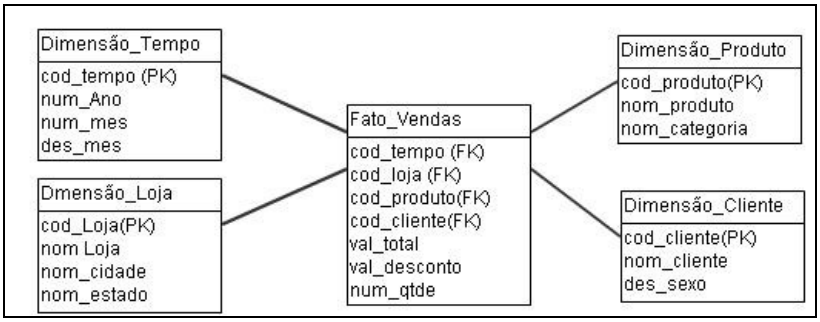

Figura 6 – Exemplo de esquema físico dimensional estrelado