Estruturando o *Business Intelligence* através do processo de *Data Warehouse*

Murillo Higor Fernandes Carvalhaes, André Luiz Alves

Pontifícia Universidade Católica de Goiás – PUC Goiás – Goiânia – GO

Pós-Graduação Qualidade e Gestão de Softwares

murillohigor@gmail.com, andre.luiz@pucgoias.edu.br

Abstract. *This article shows how information technology can support business decision making based on data from legacy systems aligned with business strategy. It presents concepts related to Business Intelligence, Data Warehouse and how the process should be built.*

Resumo. *Este artigo apresenta como a tecnologia da informação pode apoiar a empresa na tomada de decisão baseado em dados dos sistemas legados alinhados com a estratégia do negócio. É apresentado conceitos ligados ao Business Intelligence, Data Warehouse e como o processo deve ser construído.*

1. Introdução

A economia mundial hoje é baseada em informação e saber utilizá-la estrategicamente ganha eficiência e competividade. A gerência de dados ganha espaço e passa a ser parte da estratégia da empresa, administrando a informação alinhada com a estratégia de negócio da empresa. Toda de informação gerada deve ser armazenada e trata para que seja acessível de forma rápida, objetiva e simples.

A proposta do BI (*Business Intelligence*) é transformar dados em informações que possam ser usadas para ações analíticas, tomadas de decisões tático-estratégicas e até definições operacionais. A inteligência pode estar armazenada em varias formas como planilhas eletrônicas, livros, anotações em papel porém a gestão dessa forma se torna mais difícil e suscetível a erros. Dados espalhados em vários sistemas sem padronização dificultam o levantamento de informações que servirão de insumo para as reuniões estratégicas da empresa, ocorrendo por exemplo de números diferentes para o mesmo indicador.

Este artigo apresenta os principais conceitos do *Data Warehouse* e seus benefícios para o organização como apoio para o *Business Intelligence*. Na seção 2 introduz o *Data Warehouse* apresentando a motivação de seu uso e como a empresa é afetada pelo processo. Na seção 3 o processo é detalhado com a definição das etapas e como devem ser conduzidas. Em seguida na seção 4 o modelo multidimensional é exposto descrevendo seus elementos e sua implementação em um sistema de vendas. A

representação do modelo multidimensional é apresentada na seção 5 seguida da conclusão deste artigo.

2. Data Warehouse

O *DW* (*Data Warehouse*) é um conceito de como organizar os depósitos de dados corporativos, dominando as informações estratégicas para garantir respostas e ações rápidas com base nos fatos apresentados, assegurando competitividade à empresa. O *DW* não é um produto pronto a ser utilizado, ele é construído baseado nas necessidades de informação da empresa, todo o processo exige estudo e envolvimento dos executivos na definição e construção dessa base de dados.

O envolvimento pelos responsáveis pelo negócio é importante para que o *DW* seja aceito pela comunidade de negócios. De nada adianta um *DW* com milhões de dados, mas não traga os indicadores necessários para a tomada de decisão.

A construção exige a transferência e transformação dos dados coletados nas diversas bases dos sistemas transacionais, para uma base de dados independente. Essa base de dados ficará disponibilizada para os usuários, mantida por meio de processo diferenciado dos existentes para os sistemas em para os sistemas em operação transacional normais à empresa.

O *Data Warehouse* cria padrões melhorando os dados analisados de todos os sistemas, corrigindo os erros e reestruturando os dados sem afetar o sistema de operação, apresentando somente um modelo final e organizado para a análise.

As informações armazenadas são agrupadas por assuntos de interesse da empresa que são mais importantes, ou seja os principais processos da empresa, em contrate com os sistemas transacionais que são orientados a processos desenvolvidos para manter as transações realizadas diariamente.

Segundo Kimball, O *DW* deve ser um baluarte seguro que protege as informações. O tesouro das informações está guardado nele e por isso, o *DW* deve controlar de forma eficaz o acesso às informações confidenciais da empresa.

Os dados são inseridos refletindo a base histórica e não são feitas atualizações. Estarão disponíveis somente para leitura e não podem ser alterados. Eles representam resultados transacionais em determinado momento de tempo, geralmente as analises são feitas em dados dos últimos cinco, dez ou até mais anos.

Segundo Machado, os dados de um *DW* são precisos em relação ao tempo, representam resultados operacionais em determinado momento de tempo, o momento em que formam capturados. Isso implica que os dados de um *DW* não possam ser atualizados.

A data é o elemento principal do *DW*, toda sua construção é baseada em uma cronologia de fatos ocorridos ao longo do tempo.

3. Processo de *Data Warehousing*

O processo de concepção do *DW* envolve quatro etapas a extração, transformação, carga e disponibilização dos dados (Figura 1):

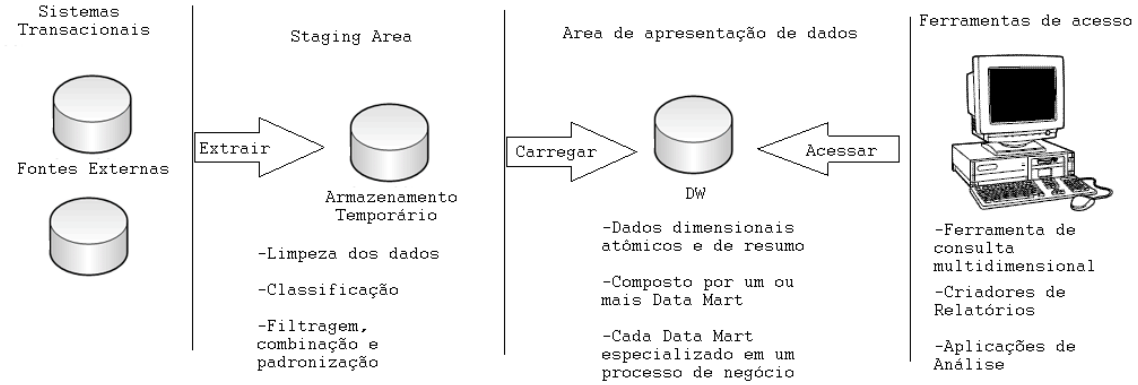

Figura 1. Etapas de construção do Data Warehouse

Fonte: Elaborado pelo autor

3.1. Extração dos dados

A primeira etapa do DW é a extração de dados de fontes externas oriundos de sistemas transacionais OLTP (*On Line Transaction Processing*) que são os sistemas onde são realizados todas as transações que serão analisadas. As origens dos dados podem ser diversas como *mainframe*, fita magnética, *web service*, base de dados em diferentes linguagens, ou até mesmo arquivo de texto.

3.2. Transformação e tratamento dos dados

Em seguida os dados podem ser inseridos em uma área armazenamento temporário chamado de "*Staging Area*", muito usual em cargas a partir de diversas bases e para evitar a concorrência com o ambiente transacional. Nessa área os dados são devidamente transformados e preparados para então inseridos no *DW*. A *Staging Area* não é uma etapa obrigatória pois os dados podem ter extraídos e transformados na memoria sem apoio de uma base de dados e inseridos diretamente no *DW*.

A utilização dessa arquitetura permite rastrear origem dos erros identificados e realizar o sincronismo por janelas estreitas de disponibilização. Exemplo supondo que o sistema A disponibiliza suas informações durante 30 minutos a partir das 3 da madrugada e o sistema B disponibiliza sua informação a partir das 5 da madrugada. O *DW* deve ser carregado quando toda a base histórica dos diversos sistemas esteja disponível para que não corra o risco de reprocessamento.

A garantia de unicidade da informação é feita durante a transformação, todos os dados devem ser padronizados, por exemplo, em um sistema podemos armazenar o sexo do cliente com 'M' ou 'F', já em outro armazenado com 1 ou 0, toma-se a decisão da padronização para que o DW tenha que lidar com formas diferentes.

As informações podem estar pulverizadas em diversas bases e durante a transformação os dados se convergem em um só registro garantindo a integridade (Figura 2).

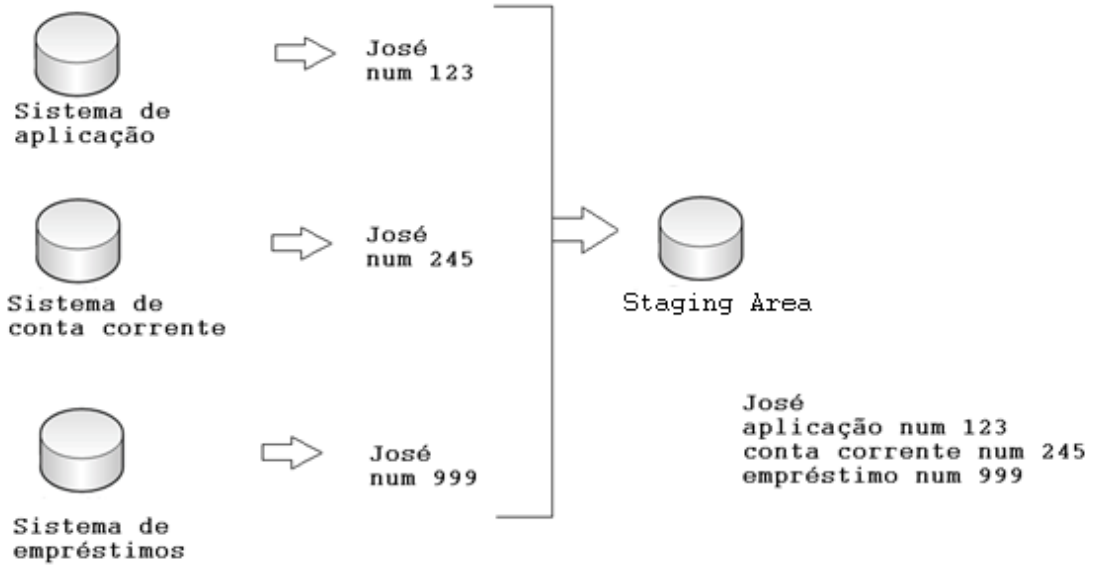

Figura 2. Unindo informações pulverizadas

Fonte: Elaborado pelo autor

3.3. Carga dos dados

Na próxima etapa após todo o tratamento os dados são carregados o *DW* que é composto de partes menores denominadas de (*DM*) *Data-Marts*. Os *DM* são as prateleiras do armazém de dados especializados em áreas especificas da empresa, por exemplo: Financeiro, Vendas, Recursos Humanos.

3.4. Acesso aos dados

Segundo Kimball, a área de apresentação de dados é o local que os dados ficam organizados, armazenados e tornam-se disponíveis para serem consultados diretamente pelos usuários, por criadores de relatórios e por outras aplicações de análise. Essa área é tudo o que a comunidade de negocio vê.

O *DW* é mantido por ferramentas do tipo OLAP (*On Line Analytical Processing*) que são capazes de realizar todas as etapas do processo de construção do *DW* através da modelagem multidimensional.

4. Modelagem Multidimensional

Nos sistemas transacionais é comumente utilizada a modelagem ER (Entidade Relacionamento), baseado na Terceira Forma Normal que reduz as redundâncias através de relacionamento entre tabelas. Com isso os dados ficam distribuídos em varias tabelas

e sua busca se dá através de consulta que requer várias junções. A principal diferença entre o modelo ER e o modelo dimensional está no nível de normalização.

Um modelo multidimensional é formado por três elementos básicos: Fatos, Dimensões e Medidas que serão descritos abaixo.

4.1. Fato

Um fato é uma medição do negócio da empresa, e representam valores numéricos através de suas medidas e é implementado em tabelas denominadas tabelas fato. Segundo Barbieri, originam-se das entidades encontradas no modelo relacional que representam ações, eventos, acontecimentos, enfim, fatos que desejamos registrar.

Como exemplo, temos abaixo um sistema de vendas, no qual o valor total de vendas de cada produto por ano é um fato (Tabela 1):

Tabela 1. Sistema de Vendas

Vendas	2012	2013	2014
Produto A	1274	967	1632
Produto B	53	70	95
Produto C	234	194	163
Produto D	562	675	841
Produto E	75	67	73

Fonte: Elaborado pelo autor

A partir dos dados apresentados podemos extrair informações analíticas, como a evolução das vendas ano a ano, e traçar estratégia de vendas de acordo com as diretrizes da empresa. Identificar quais os possíveis motivos de baixas de um ano para o outro etc. Os fatos variam ao longo do tempo (dimensão tempo), possuem valores numéricos (medidas) e seu histórico pode ser mantido e cresce ao longo do tempo para cada item (dimensão produto).

4.2. Dimensão

As dimensões são elementos que determinam o contexto de um assunto de negócio. São as possíveis formas de visualizar os dados, ou seja, por mês, por pais, etc.

Segundo Barbieri, as dimensões possuem múltiplas colunas de informação, algumas das quais representam a sua hierarquia. Apresentam sempre uma chave primaria que lhes confere unicidade, chave esta que participa da tabela fato como parte da sua chave múltipla.

No exemplo do sistema de vendas apresentado temos duas dimensões, tempo (por ano) e produto (por produto). Na dimensão produto podemos temos o detalhamento das características do produto, como a cor, peso etc.

4.3. Medida

Segundo Machado, as Medidas são os atributos numéricos que representam um fato. Uma medida é determinada pela combinação das dimensões que participam de um fato, e estão localizadas como atributos de um fato.

No exemplo do sistema de vendas apresentado os valores de venda são as medidas. O relacionamento entre as entidades fato e dimensão para o sistema de vendas pode ser visualizado na Figura 3:

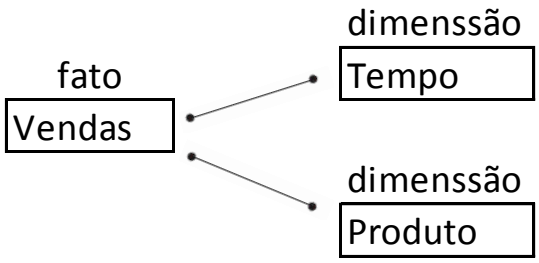

Figura 3. Relacionamento Fato Dimensão Vendas

Fonte: Elaborado pelo autor

A implementação da tabela fato deve conter as chaves para todas as dimensões ligadas a ela e o valor de suas medidas (Tabela 2), geralmente o identificador do fato é formado pela composição de todas as chaves estrangeiras das dimensões:

Tabela 2. Implementação Fato Vendas

id_fk_produto	id_fk_tempo	nu_vendas
1	1	1274
2	1	53
3	1	234
4	1	562
5	1	75
1	2	967
2	2	70
3	2	194
4	2	675
5	2	67
1	3	1632
2	3	95
3	3	163
4	3	841
5	3	73

Fonte: Elaborado pelo autor

A implementação da dimensão é composta por uma chave única em seguida das características da dimensão que devem ser únicas (Tabela 3).

Tabela 3. Implementação Dimensão Tempo

id_tempo	nu_ano
1	2012
2	2013
3	2014

Fonte: Elaborado pelo autor

Pode ocorre que um item da dimensão tenha uma variação em uma ou mais características como, por exemplo, mudança da cor. Como exemplo na Tabela 4 o produto A tem a cor vermelha, mas em um determinado momento passou a ser preta, o registro vermelho id_produto igual a 1 não pode ser atualizado para a cor preta, o correto é criar outro item na dimensão variando somente a cor para preta. A justificativa para isso é manter os dados históricos, caso seja feita a atualização da cor no produto com id_produto igual a um, o fato venda de produtos A na cor vermelha passa a não existir no *DW* tendo impacto na analise dos dados.

Tabela 4. Implementação Dimensão Produto

id_produto	no_produto	no_cor	no_peso
1	Produto A	Vermelha	64
2	Produto B	Preta	234
3	Produto C	Amarela	110
4	Produto D	Azul	75
5	Produto A	Preta	64

Fonte: Elaborado pelo autor

O fato é uma venda realizada, nesse modelo temos o tempo e o produto que fazem parte da venda. Poderíamos ter outras dimensões como Cliente, Vendedor, Filial etc. O *DW* terá parte dos dados do sistema transacional. O sistema transacional pode armazenar as informações de vendas na hora exata que aconteceu porém nesse caso temos interesse na evolução por ano. Essa decisão é baseada na necessidade do negócio que poderia ser por trimestre, ou mês ou até mesmo por dia.

5. Representação do Modelo Multidimensional

Segundo Machado, o modelo mais popular para visualizar um modelo dimensional é o desenho do cubo. Quando o modelo dimensional tem mais de três dimensões é chamado de hipercubo. As combinações das dimensões levam a outro cubo, até que se chegue à medida buscada. A metáfora denominada CUBO é apenas uma aproximação da forma como os dados estão organizados, e não a rela expressão de uma realidade.

Vamos adicionar ao nosso modelo a dimensão filial, no qual identifica em qual unidade da empresa ocorreu à venda. Na figura abaixo temos a representação do cubo contendo a dimensão produto, tempo e filial. Ao buscar, por exemplo, pela filial DF no ano 2012 o produto A encontramos a media 1274, ou seja, o fato é que na filial DF no ano de 2012 o valor total de vendas do produto A foi de 1274 reais.

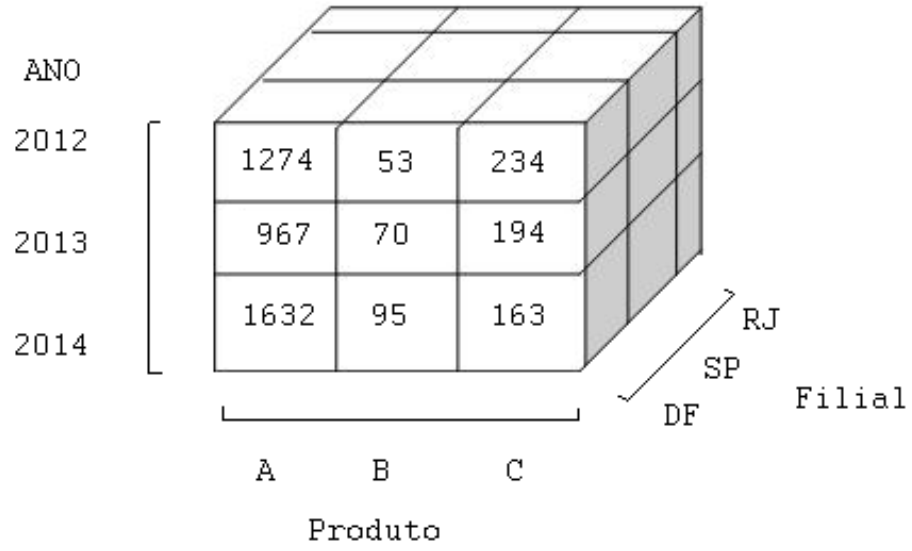

Figura 4. Cubo

Fonte: Elaborado pelo autor

Modificamos a dimensão tempo e inserimos a informação dos trimestres do ano. No momento de gerarmos o cubo fazemos o agrupamento por trimestre de cada ano. A criação de membros na dimensão gera a hierarquia. A hierarquia de uma dimensão é uma classificação de dados dentro de uma dimensão. Neste caso o primeiro nível é o ano e o segundo o trimestre. A quantidade de membros na hierarquia da dimensão depende da necessidade, na dimensão tempo os níveis abaixo mais usuais são: mês, dia, hora, minuto, segundo.

Ao analisarmos o cubo de vendas queremos visualizar as vendas por trimestre no ano de 2012. Faremos o *Drill Down*, ou seja, aumentamos o nível de detalhe da informação. Os dados serão apresentados por trimestre do ano de 2012. O *Drill Down* é uma operação do modelo multidimensional, a capacidade de detalhar os dados que estão sumarizados.

O movimento de sumarização, ou diminuição no nível de detalhes é chamado de *Roll up* no caso partimos da visualização em trimestres do ano para visualização anual. Estas operações são executadas por uma ferramentas de analise multidimensional, sempre que a modelagem do cubo permitir. No exemplo acima as operações de *Drill Down* e *Roll Up* só foram possíveis depois que inserimos um novo membro na dimensão tempo criando a hierarquia que define os níveis da dimensão.

O nível de detalhes dos dados é um dos aspectos mais críticos do *DW*. Quanto mais detalhes existir, maior será a quantidade de dados inseridos no *DW*, impactando no armazenamento e no desempenho nas consultas a base.

Em contra partida se o nível de detalhes for pequeno pode não atender ao negócio em sua tomada de decisão. A construção do *DW* não existe uma formula pronto, depende sempre da necessidade em que vamos definindo quais os fatos a serem inseridos, as dimensões e seus membros e o que deve ser medido pelo cubo. Cada cubo contem um fato, e o conjunto de vários cubos forma o *DW*.

5.1. Modelo *Star* (Estrela)

O modelo estrela é o mais usual para representar o modelo de dados multidimensional, nele a tabela fato é fixada no centro e as tabelas dimensões são dispostas ao seu redor em ligação direta dimensão fato, formando uma estrela. Nesse modelo a quantidade de dimensões determina a quantidade de pontas que terá a estrela. O modelo estrela é a estrutura básica de um modelo de dados multidimensional.

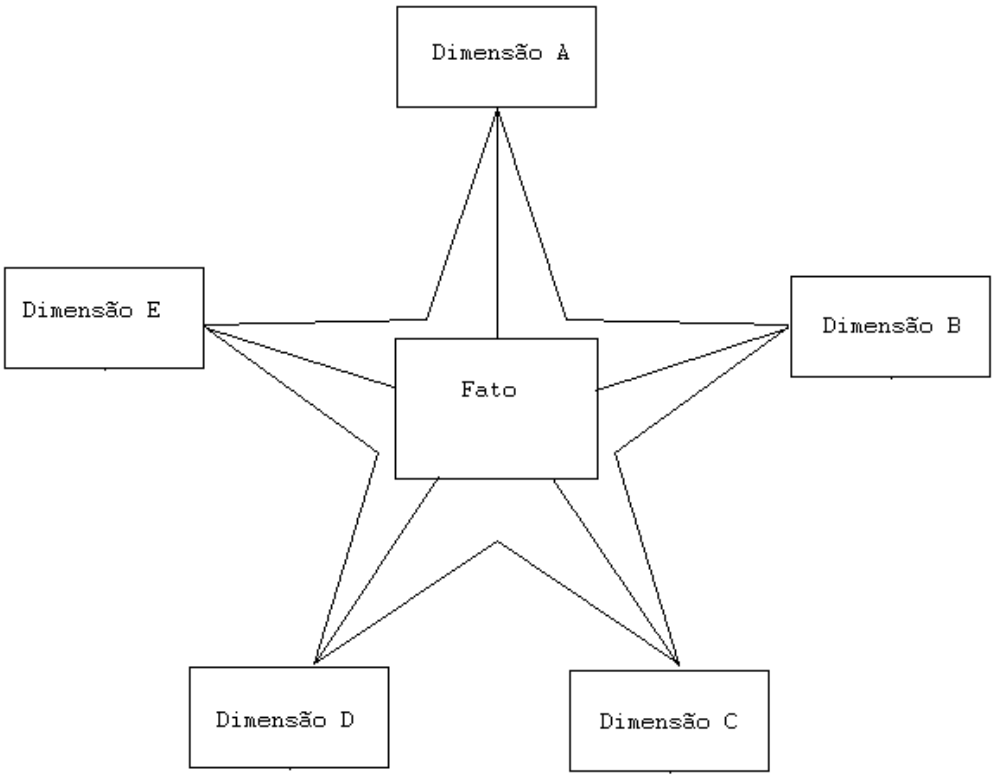

Figura 5. Modelo Estrela

Fonte: Elaborado pelo autor

5.2. Modelo *Snowflake* (floco de neve)

Outro modelo utilizado é o floco de neve que parte do modelo estrela e decompõe uma ou mais dimensões que possuem hierarquias entre seus membros aplicando a terceira forma normal sobre a dimensão (Figura 6).

A vantagem desse modelo é a redução da redundancia na tabela dimensão e diminuindo o espaço alocado no armazenamento. Em contrapartida aumenta os *joins* nas consultas à dimensão, podendo prejudicar o desempenho do *DW*.

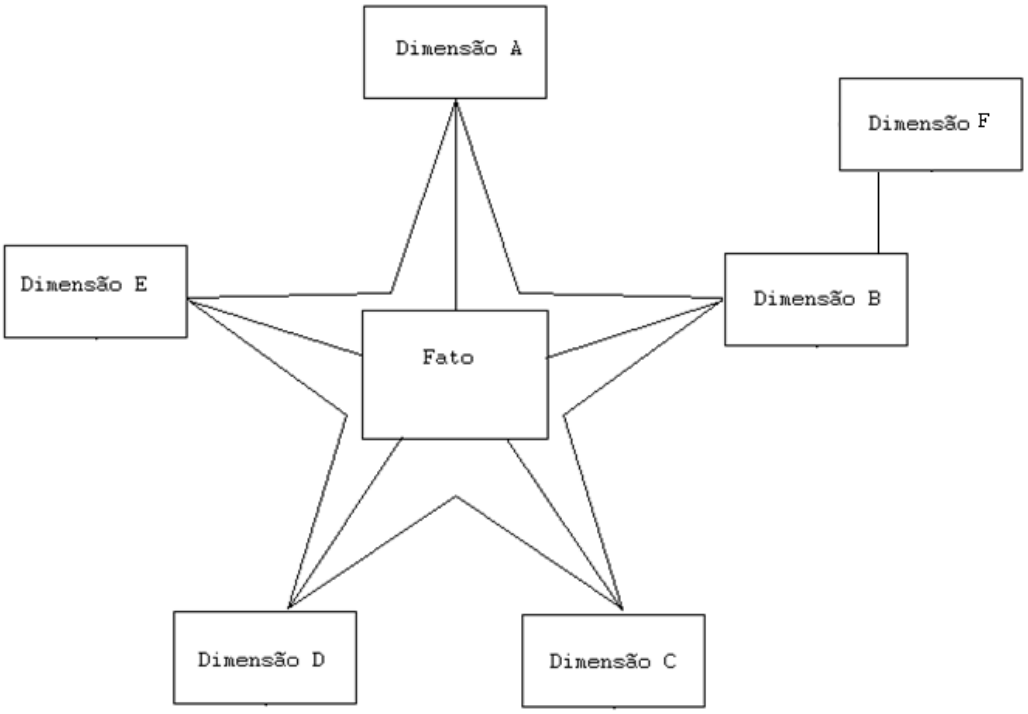

Figura 6. Modelo floco de neve

Fonte: Elaborado pelo autor

Na tabela 4 foi adicionado o "produto A" com o atributo cor preta que se repete em dois registros distintos. Se colocarmos as cores em outra tabela e referenciarmos a ela dentro da dimensão Produto temos a aplicação do modelo floco de neve (Figura 7):

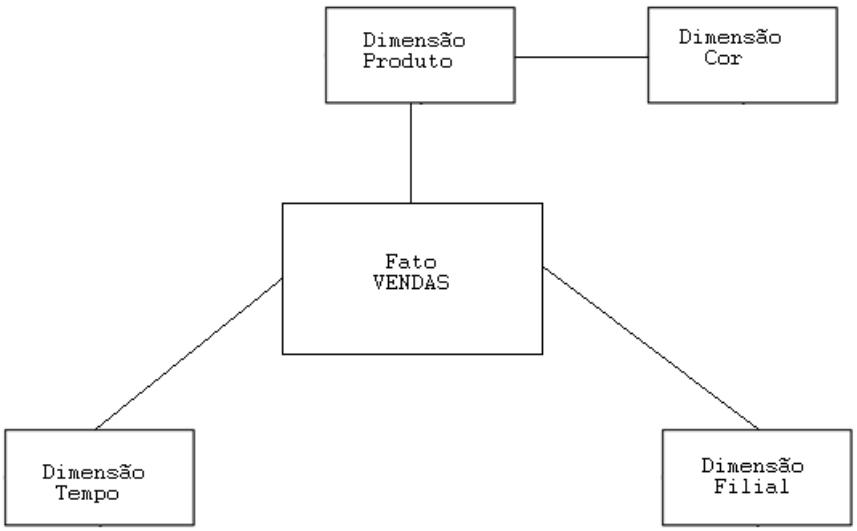

Figura 7. Sistema de Vendas Modelado em floco de neve

Fonte: Elaborado pelo autor

6. Conclusão

Neste artigo demonstramos que muitas vezes as empresas tomam decisão sem embasamento devido à falta de estrutura dos dados históricos e a dificuldade de acessar tais dados de maneira consolidada. Vimos que não adianta somente estruturar os dados se eles não estiverem em uma linguagem de fácil entendimento e realmente tenham as informações relevantes para o negócio. Muitas das vezes o fracasso de um *DW* se da pelo falta de apoio da área de negócio e as informações disponibilizadas não são relevante para os administradores. Ou seja, o projeto não foi alinhado com as necessidades de quem for utilizado.

Trata-se de uma nova forma de modelar os dados, demonstramos os principais componentes da modelagem multidimensional e como é estruturada. A implantação de um processo de *Data Warehousing* não é uma tarefa trivial, exige mais do que apenas investimento financeiro, exige alto grau de comprometimento da diretoria seja na sua elaboração como no uso.

Foi apresentado a aplicação do processo em um sistema de vendas de produtos e vimos como podemos estruturar os dados para analise da evolução de vendas ano a ano. O *DW* em si não é não implica necessariamente em melhores decisões, ele gera a informação para que os administradores se baseiem suas decisões. Cabe à empresa identificar as necessidades de informação para que o *DW* seja moldado para atender tal fim.

7. Referências

Armazém de Dados, https://pt.wikipedia.org/wiki/Armazém_de_dados, acesso em: 02/07/2015.

Barbieri, Carlos "BI2 – Business Intelligence: modelagem e qualidade". Rio de Janeiro: Editora Elsvier 2011.

Data Warehousing Concepts, https://docs.oracle.com/cd/B10500_01/server.920/a96520/concept.htm , acesso em: 28/07/2015.

Machado, Felipe Nery Rodrigues "Tecnologia e Projeto de Data Warehouse: uma visão multidimensional". 5º Edição Revisada e Atualizada. São Paulo. Editora Érica Ltda 2010.

Modelagem Multidimensional, https://pt.wikipedia.org/wiki/Modelagem_dimensional, acesso em: 01/07/2015.

Ralph Kimball, MargyRoss "The Data Warehouse Toolkit – O guia complete para modelagem multidimensional". Tradução de Ana Beatriz Tavarez, Daniela Lacerda. Rio de Janeiro, Editora Campus 2002.